A Reference Sequence for *Blumeria graminis* f. sp. *tritici* (Wheat Powdery Mildew) and its Application for Comparative and Evolutionary Genomics

---

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Simone Oberhänsli

von

Kemmental TG

Promotionskomitee

Prof. Dr. Beat Keller (Vorsitz und Leitung der Dissertation)
PD Dr. Thomas Wicker
Prof. Dr. Robert Dudler

Zürich, 2013

Se que hay en tus ojos con solo mirar
que estas cansado de andar y de andar
y caminar, girando siempre en un lugar

Se que las ventanas se pueden abrir
cambiar el aire depende de ti
te ayudará, vale la pena una vez mas

Saber que se puede, querer que se pueda
Quitarse los miedos, sacarlos afuera
pintarse la cara color esperanza
tentar al futuro con el corazón

Es mejor perderse que nunca embarcar
mejor tentarse a dejar de intentar
aunque ya ves que no es tan facil empezar
Se que lo imposible se puede lograr
que la tristeza algun día se irá
y asi será, la vida cambia y cambiará

Sentirás que el alma vuela
por cantar una vez mas

Saber que se puede querer que se pueda
quitarse los miedos, sacarlos afuera
pintarse la cara color esperanza
tentar al futuro con el corazón

Vale más poder brillar
que solo buscar ver el sol

Pintarse la cara color esperanza
tentar al futuro con el corazón

"Color Esperanza", a song by Diego Torres

Contents

Summary

Powdery mildew is one of the most important cereal diseases. It is caused by fungi of the species *Blumeria graminis*, which have an obligate-biotrophic lifestyle and are specific for a single host plant species. The topic of this thesis is the molecular analysis of the genome sequence of *Blumeria graminis* forma specialis *tritici* (*B.g. tritici*), the powdery mildew pathogen of wheat, with regard to its biology as a plant pathogen.

Powdery mildew of wheat and barley are caused by *B.g. tritici* and *B.g. hordei*, respectively. In a pilot study, a comparative analysis of two orthologous loci in the genomes of the two fungi revealed that the orthologous genes are well conserved and synthenic, whereas the intergenic regions are highly diverse and massively populated with transposable elements (TEs). A divergence time estimate based on the sequence alignments indicates that *B.g. tritici* and *B.g. hordei* have co-evolved with their hosts.

The major part of this work reports on the reference sequence of *B.g. tritici* (isolate 96224) and it's bioinformatic analysis. *B.g. tritici* was found to have a large with a high content of repetitive DNA and TEs. To get an overview about the TE population in the *B.g. tritici* genome, we produced a repeat library that contains the 56 most abundant TEs. The library was also used to annotate TEs in the genome sequence. Similar to what was found for other biotrophic fungi, *B.g. tritici* lacks a number of genes which code for proteins that are involved in the degredation of cell wall components, carbohydrate transport, nitrate and sulfur metabolism or the production of secondary metabolites. Based on comparative analysis with *B.g. hordei*, we identified about 600 candidate effector genes that could play a role during the infection process or the determination of host specificity.

In addition to the Swiss isolate 96224, we re-sequenced the genomes of three wheat powdery mildew isolates from UK, Israel and Switzerland. Compared to the reference (isolate 96224), the three re-sequenced isolates differ in the absence of certain genes, most of them are effector candidates. Apparently, there is a selective pressure for loosing these genes which makes them candidates for determinants of race specific interactions. Single nucleotide polymorphisms present in the genome of the three isolates compared to the reference were found to be unevenly distributed, which leads to a mosaic structure of these genomes consists of younger and more ancient regions. These highly diverse haplogroups have already existed prior to the domestication of wheat. We hypothesize that the occurence of bread wheat as a new host for *B.g. tritici* did not lead to a dramatic loss of genetic diversity in the genome, and that the highly diverse haplotype pool provides a large genetic potential for pathogen variation facilitating its ready adaptation to new host species.

Zusammenfassung

Mehltau ist weltweit eine der verheerendsten Getreidekrankheiten. Verursacht wird sie von Pilzen der Art *Blumeria graminis*, welche eine obligat-biotrophe Lebensweise haben und wirtsspezifisch, also spezialisiert auf eine Pflanzenart sind. Diese Arbeit befasst sich mit der molekularen Analyse der Genomsequenz von *Blumeria graminis* forma specialis *tritici* (*B.g. tritici*), dem Verursacher von Mehltau bei Weizen, hinsichtlich seiner Biologie als Getreidepathogen.
Mehltau bei Weizen und Gerste wird von zwei unterschiedlichen Pilzen verursacht, welche jeweils spezifisch für ihre Wirtspflanze sind. Ein Sequenzvergleich zweier orthologer Loci aus den Genomen der beiden Pilze hat gezeigt, dass die orthologen Gene synthenisch und stark konserviert sind, während sich die intergenen Regionen grundlegend unterscheiden und einen sehr hohen Anteil an Transposons enthalten. Der Sequenzvergleich ermöglichte ausserdem eine Schätzung des Zeitpunkts der Trennung der beiden Pathogene von ihrem letzten gemeinsamen Vorfahren, was auf eine Co-evolution der beiden Pathogene mit ihren jeweiligen Wirten hindeutet.
Der Hauptteil der Arbeit widmet sich der vollständigen Sequenzierung des Weizenmehltau-Genoms (Isolat 96224) und seiner bioinformatischen Analyse. Mehltaupathogene haben, verglichen mit dem anderen Pilzen, ein relativ grosses Genom, welches reich bevölkert ist mit repetititven Elementen, sogenannten Transposons. Mit dem Ziel einen Überblick über die Transposonpopulation im Weizenmehltaugenom zu erhalten, haben wir eine Transposon-Datenbank mit den 56 häufigsten repetitiven Elemente, welche im Weizenmehltaugenom vorkommen, erstellt. Diese Datenbank wurde benützt um Transposons in der Genomsequenz zu annotieren. Wie schon bei anderen biotrophen Pilzen gefunden wurde, fehlen im *B.g. tritici* Genom bestimmte Gene, zum Beispiel solche, die für Proteine kodieren, die in den Abbau von Zellwandkomponenten, Kohlenhydrattransport, Nitrat- und Schwefelmetabolismus oder Produktion von Sekundärmetaboliten involvert sind. Mittels vergleichender Studien mit *B.g. hordei* konnten etwa 600 Gene identifiziert werden, welchen eine tragende Rolle während dem Infektionsprozess oder der Festlegung der Wirtspezifität zugeschrieben wird (sogenannte Effektoren).
Zusätzlich zum Schweizer Isolat 96224 wurden noch drei weitere Mehltauisolate sequenziert, welche in Grossbritannien, Israel und der Schweiz gesammelt wurden. Im Vergleich zur Referenz, dem Isolat 96224, fehlen den drei zusätzlich sequenzierten Isolaten einzelne Gene, von denen fast alle Effektorkandidaten sind. Diese Gene, welche offensichtlich einer Selektion bezüglich deren Verlust unterliegen, könnten als entscheidende Faktoren der Rassenspezifität der Isolate wirken. Ausserdem unterscheiden sich die drei re-sequenzierten Isolate im Vergleich zur Referenz auch in einzelnen Basen, sogenannten SNPs (single nucleotide polymorphisms). Diese sind ungleich im Genom verteilt und führen zu einer Mosaikstruktur von "älteren" uns "jüngeren" Bereichen. Diese höchst verschiedenartigen Haplogruppen haben schon vor der Domestikation von Weizen existiert. Wir vermuten, dass die Enstehung von Brotweizen als neue

Wirtpflanze nicht zu einer Reduktion der genetischen Vielfalt im *B.g. tritici* Genom geführt hat, und dass die hohe Vielfalt an Haplotypen ein grosses genetisches Potenzial für Pathogenvariablilität darstellt, welches die schnelle Anpassung an neue Wirtspflanzen begünstigt.

CHAPTER 1

Introduction

1.1 The powdery mildew disease

Powdery mildew is a fungal disease that affects nearly 10,000 species of angiosperms, among them many economically important cultivated plants such as ornamental plants, fruit trees and cereals. Mildew pathogens are ascomycete fungi of the order Erysiphales, which includes approximately 820 species (Braun, 2011). Cereal powdery mildew is caused by *Blumeria graminis* and is only found on wild grasses and cultivated cereals of the family Poaceae. Economically most important are powdery mildew of wheat and barley, which cause significant yield loss in regions with maritime or semi-continental climates with high rain fall in many parts of the world, for example in China, Russia and Europe. The main measures of disease control are breeding for resistant varieties and application of fungicides.

Blumeria graminis (former name *Erysiphe graminis*) are obligate biotrophic fungi which can only grow on living cells of their host plants. Airborne spores start to germinate as soon as they land on the plant leaf and form an appressorial hook by which they penetrate the host epidermal cell wall (Figure 1.1a, Zhang *et al.* 2005). Once the fungus has gained access to the cell lumen, a finger-like structure called haustorium develops in the intercellular space (Figure 1.1a and b). This organ is enclosed within the extrahaustorial matrix, a membrane composed of modified host cell plasmalemma, and serves as an interface for nutrient uptake from the plant to the fungus (Zhang *et al.*, 2005). Once haustorial nutrient supply has been established, secondary hyphae will spread out on the leaf surface to further invade neighbouring cells. Clonally produced conidiospores stored in conidiophores on secondary hyphae on the plant surface lead to further dispersal of the disease and give the fungus the powdery-like appearance from which the name comes from (Figure 1.2c and d).
Besides clonal reproduction, which takes about 14 days, *Blumeria graminis* also has a sexual phase in its life cycle (Figure 1.3). It requires hyphal fusion of two fungal colonies of different mating type and usually occurs close to the end of the growing season when the leaves of the host plant dry out. Sexual recombination takes place in so-called chaesmothecia and results in haploid ascospores which remain inside the chaesmothecia for overwintering (Glawe, 2008). In spring, the ascopores are ejected and new infection cycles start.

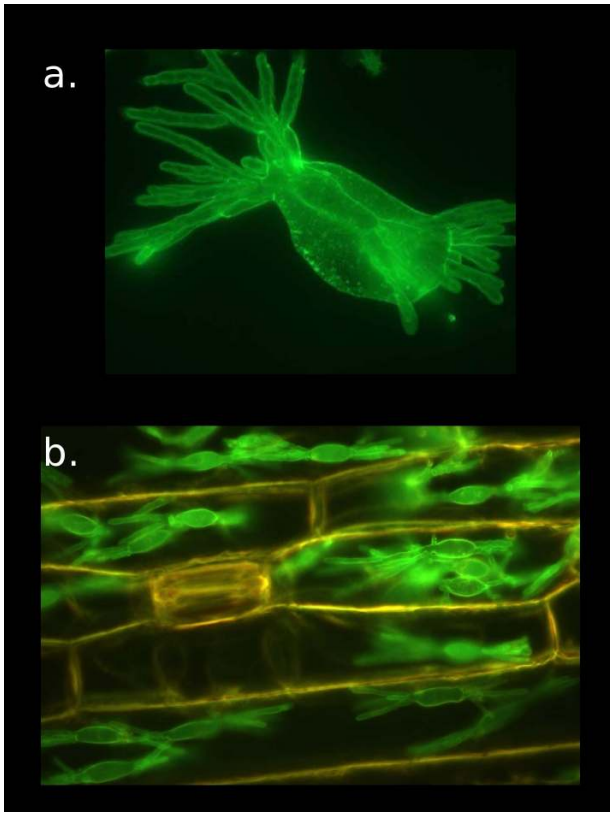

Figure 1.1. Powdery mildew infection structures. **a.** Purified haustorium and host perihaustorial membrane. **b.** Barley epidermal cells (yellow) with haustorial structures (green). (Pictures: P. Spanu)

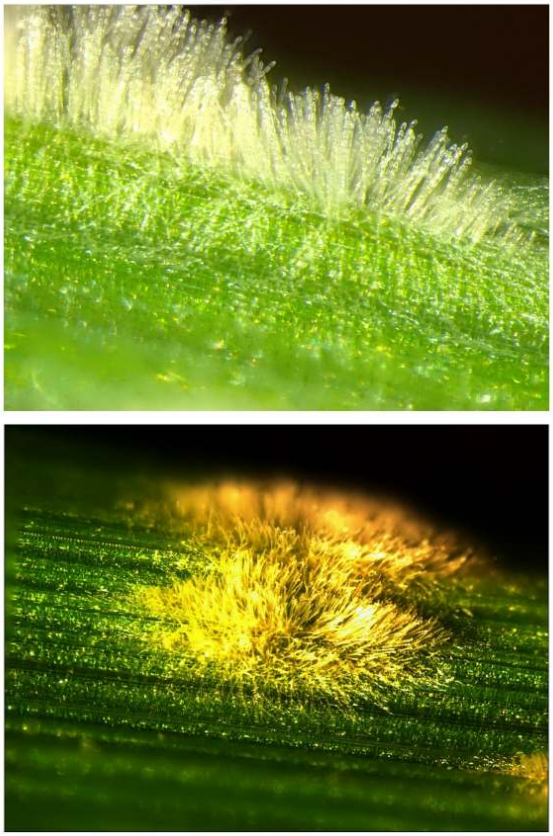

Figure 1.2. Powdery mildew colonies on barley leaves. (Pictures: P. Spanu)

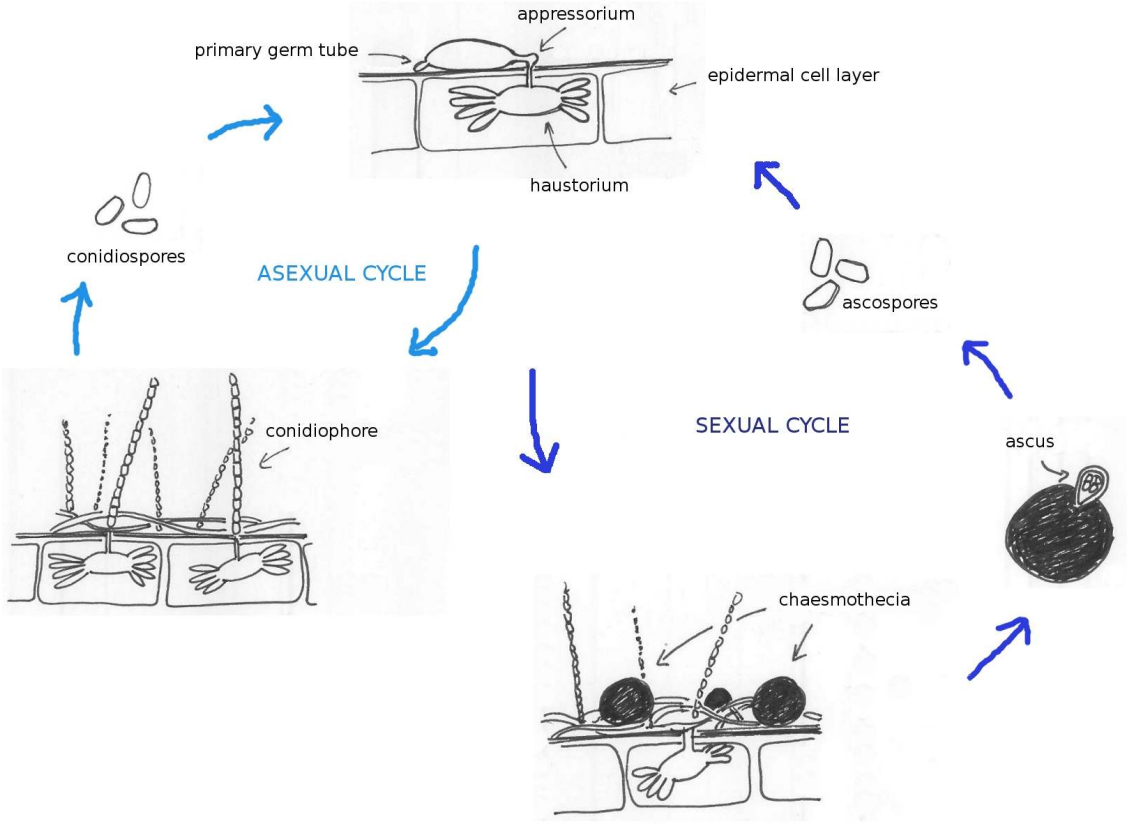

Figure 1.3. Sexual and asexual life cycle of cereal powdery mildews. Modified after G.L. Schumann from "Plant Diseases: Their Biology and Social Impact", APS Press

1.2 Host specificity of *Blumeria graminis*

Blumeria graminis fungi are host specific pathogens that in general infect only one specific plant species. Seven *Blumeria graminis formae speciales* were described by Marchal (1902) based on host specialisation. This number was extended to eight by Oku *et al.* (1985) (Table 1.1). All *Blumeria graminis formae speciales* have their characteristic adaptation, but host specificity is not complete. Studies in mostly older literature (reviewed in Hiura, 1978) report on infection experiments where *formae speciales* are not limited in growth to a single genus (Eshed and Wahl, 1970; Hardison, 1944; Mühle and Frauenstein, 1962a,b, 1963). Under natural conditions however, powdery mildew of barley, wheat and rye favor their own host and rarely infect other crop species (Eshed and Wahl, 1975; Mühle and Frauenstein, 1970).
The genetic basis of powdery mildew *formae speciales*-host specificity is not yet well known. Two studies conclude that it follows the gene-for-gene theory, meaning that incompatibility of pathogen and host is conferred by a pair of matching genes present in the fungus and the plant, respectively (Flor, 1971). Tosa (1989) used F1 and F2 hybrids of a cross between a *B.g. tritici* isolate (Tk-1) with an *B.g. agropyri isolate* (Ak-1) to conduct infection experiments on a few wheat cultivars. The segregation ratios of compatibility and incompatibility in these experiments could be well explained with the gene-for-gene theory. Tosa hypothesizes that a *forma specialis* carries many avirulence genes that correspond to resistance genes of more than one plant genus, and in accordance with Hiura (1978), speculates that host specialization developed gradually and therefore the present *formae speciales* still share many loci for pathogenicity. Oku and Tsuchizaki (1993) report that the results of infection experiments with hybrids of *B.g secalis* and *B.g. tritici* on wheat lines with and without rye resistance genes can be explained by the gene-for-gene hypothesis as well. The authors conclude that hybridization between *B.g secalis* and *B.g. tritici*, if it actually happens under field conditions, could lead to compatibility of *B.g. tritici* with wheat-rye translocation lines. Interestingly, it was just recently discovered that *B.g. tritici* isolates can now infect triticale, an artificial hybrid species of wheat and rye only grown in agriculture for about 50 years (Walker *et al.*, 2010; Troch *et al.*, 2012). Walker *et al.* (2010) however conclude that adaptation of *B.g. tritici* to the new host triticale by host range expansion is more likely than a hybridization between *B.g. tritici* and *B.g. secalis*.

Hiura (1978) studied compatibility between *formae speciales hordei*, *tritici*, *agropyri*, *secalis*, *avenae* and *poae*. Hybridization of two *formae speciales* can take place on any of the host cultivars, because plant cells which are already infected by the appropriate *forma specialis* are also accessible for other *formae speciales* (Olesen *et al.*, 2003). According to Hiura's observations, hybridization between *forma specialis agropyri*, *tritici* and *secalis* results in normal formation of chaesmothecia and ascospores, if two different mating types are present. When *B.g. hordei* and *forma specialis agropyri*, *tritici* or *secalis* were hybridized, normal chaesmothecia development was observed, but the asci contained only very few ascospores. *Forma specialis avenae* and *poae* were both not compatible with either of the other *formae speciales*, which suggests that there might be several levels of reproductive isolation among *formae speciales*.

Table 1.1. *Blumeria graminis formae speciales* and their hosts

Pathogen	Host		
	Common Name	Genus	Tribe
Blumeria graminis f.sp. *tritici*	wheat	*Triticum*	Triticeae
Blumeria graminis f.sp. *hordei*	barley	*Hordeum*	Triticeae
Blumeria graminis f.sp. *secalis*	rye	*Secale*	Triticeae
Blumeria graminis f.sp. *agropyri*	wheatgrass	*Agropyron*	Triticeae
Blumeria graminis f.sp. *avenae*	oat	*Avena*	Avenae
Blumeria graminis f.sp. *bromi*		*Bromus*	Bromeae
Blumeria graminis f.sp. *poae*		*Poa*	Poeae
Blumeria graminis f.sp. *lollii*	ryegrass	*Lolium*	Poeae

1.3 Triticeae crops as host species of *Blumeria graminis*

The tribe Triticeae includes over 400 species, among them the major global food crops wheat, barley and rye as well as grasses which are important for livestock forage and soil stabilization. The domestication of wheat and barley from wild ancestors in the fertile crecent about 10,000 years ago (Zohary *et al.*, 2012) substantially contributed to the rise of sedentary societies during the neolithic revolution and therefore had a significant impact on human culture. Today, wheat is after corn and rice the third most produced crop in the world (data 2010[1]) and the leading source of vegetable protein in human nutrition.

The *Triticeae* tribe includes species with diploid, allotetrapoloid or allohexaploid genomes. Some of the cultivated species have quite complex genome structures which mainly derived from hybridisation between wild grasses. Bread wheat, *Triticum aestivum*, represents 95% of today's world wheat production, the remaining 5% is *Triticum durum* which is used to make pasta. Both evolved from tetraploid wild emmer (*Triticum turgidum subsp. dicoccoides*, AABB), which resulted from hybridization of wild diploid wheat (*Triticum urartu*, AA genome) and a close ancestor of the goat grass *Aegilops speltoides* (BB genome) 300,000-500,000 years ago in the fertile crescent of the Near East (reviewed in Peng *et al.* (2011)). Domestication of wild emmer about 10,000 years ago led first to a domesticated form of emmer (*Triticum turgidum subsp. dicoccum*, AABB), and then to *Triticum durum*, the pasta wheat. Early spelt wheat (*Triticum spelta*, AABBDD), which further evolved into modern bread wheat (*Triticum aestivum*, AABBDD), was the outcome of a spontaneous hybridization between domesticated emmer and *Aegilops tauschii* (DD genome) 9,000 years ago.

[1] `http://faostat.fao.org`

1.4 Molecular basis of plant infection by fungal pathogens

Fungal pathogens use so-called effector molecules to promote infection and manipulate the plant's innate immunity system with the aim of suppressing resistance responses triggered by pathogen-associated molecular patterns (PAMPs) such as chitin (Jones and Dangl, 2006). These effectors however can become avirulence factors (AVR) which themselves trigger defence responses if they are recognized by plant resistance proteins. An example is resistance of flax to infection of flax rust *Melampsora lini*, a biotrophic pathogen: Fungal avirulence proteins AVRL567 are recognized by plant resistance proteins L567 directly in a gene-for-gene specific manner (Dodds *et al.*, 2006). The result is hypersensitive cell death response (HR) of the infected cells which leads to the death of the fungus.
The 'gene-for-gene' relationship between R- and AVR-gene causes selective pressure on the pathogen to alter or even loose AVR genes to avoid recognition (Jones and Dangl, 2006). Over time, this unintentional 'liaison' between the host and its parasite results in an evolutionary "arms-race", which forces both plant and parasite to constantly evolve in order to avoid or accomplish infection, respectively. Consequently, sophisticated survival strategies can be found on both sides. *Stagonospora nodorum* for example even utilizes R-gene mediated resistance for its own purposes: So-called necrotrophic effectors, host-specific toxins SnToxN, interact with plant sensitivity proteins (SnnN). This induces necrosis of the infected cells and the fungus, a necrotroph, can feed from the dead tissue. In this case, the R-gene of the plant, which resembles a typical biotrophic resistance gene, acts as a susceptibility gene and thus, the interaction was described as the inverse gene-for-gene model (Oliver *et al.*, 2012).

Race-specific powdery mildew resistance in cereals is assumed to follow the gene-for-gene concept, and a number of R-genes have been cloned during the last decade. In barley, the genes of the MLA complex encode allelic CC-NB-LRR receptors which confer isolate specific resistance against *B.g. hordei* (Halterman *et al.*, 2001; Seeholzer *et al.*, 2010). The wild type *Mlo* genes, in contrast, act as susceptibility factors while homozygous mutant (*mlo*) alleles of the gene confer broad-spectrum disease resistance to *B.g. hordei* (Jørgensen, 1992; Büschges *et al.*, 1997). In wheat, several powdery mildew resistance gene loci have been described, but so far only the *Pm3* allelic series (Yahiaoui *et al.*, 2004, 2006; Bhullar *et al.*, 2010) and a key member of the *Pm21* locus are cloned (Cao *et al.*, 2011).
Efforts to identify and clone powdery mildew avirulence genes have so far been moderately successful. The *B.g. hordei* AVRa10 and AVRk1 genes, which were cloned by a map-based cloning approach in 2006 (Ridout *et al.*, 2006), were proposed to be avirulence genes recognised by *Mla10* and *Mlk1*, respectively, and were shown to enhance infection in susceptible barley cultivars. These genes are paralogs which code for short proteins which have no homology to other genes in public databases. The homology to a family of LINE (long interspersed nuclear elements) transposable elements (TE) (Sacristán *et al.*, 2009), the vast amount of paralogs in the *B.g. hordei* genome (>1,350, (Spanu *et al.*, 2010) and the unknown biological function make them enigmatic proteins.
Since the release of the *B.g. hordei* genome sequence, bioinformatic screens have yielded 491 candidate secreted effector proteins (CSEPs) (Spanu *et al.*, 2010; Pedersen *et al.*, 2012), small proteins without transmembrane domain and no homology outside the Erysiphales. The delivery

mechanism of powdery mildew effectors into the plant cell is not well understood. Unlike in bacteria, which use a type-III secretion system to inject effectors into the host cell, transport of fungal effectors is most likely mediated by secretion signals. In many oomycetes effectors, an N-terminal RXLR-motif (Arg-X-Leu-Arg, where X is any amino acid) was found to be conserved and shown to be important for translocation of the effectors into plant cells (Whisson *et al.*, 2007; Dou *et al.*, 2008). Of the 491 *B.g. hordei* CSEPs studied by Pedersen *et al.* (2012), 63% contain an N-terminal YXC-motif (Tyr/Phe/Trp-X-Cys, where X is any amino acid), a motive which was described as possible secretion signal for *B.g. hordei* effector candidates (Godfrey *et al.*, 2010).

1.5 Fungal genomics as a tool to improve disease control

> *"When I began my career, I never imagined that someday I could simply look up a gene's coding sequence; find its orthologs in other organisms; and order, from a service organization, a mutation to my specification for an experiment to reveal gene function. Yet this is now our world, the direct result of a collective agreement to make genomic sequencing a priority in the last decades of the 20th century. It was a very good decision."* David Botstein, Science essay, Feb 2011.

Understanding a fungus' lifestyle and its infection strategies can significantly improve pest management. Eradication of barberry (*Berberis vulgaris*), the alternate host of rust where it sexually reproduces, together with crop rotation in areas of primary inoculum, has helped to decrease rust disease in China. The most important measure of disease control however is breeding for resistant varieties which was done successfully for the past decades. For example, the discovery of *S. nodorum* host specific toxins has resulted in breeding-programs against the corresponding susceptibility genes in wheat. However, pathogens evolve along with their hosts and many resistance genes are eventually overcome. The disastrous consequences of a break-down of a widely used resistance gene (*Sr31*) were demonstrated by Ug99, the *Puccinia graminis triticina* race TTKS which appeared in 1998 in Uganda and has since migrated rapidly over eastern Africa because most cultivars grown in this area are highly susceptible (Singh *et al.*, 2011).

Advances in genomics have tremendously improved our knowledge on the structure of fungal and oomycete genomes: The 240 Mb genome of *Phytophthora infestans* is highly repetitive, and massive proliferation of TEs is assumed to have caused the genome expansion. *Phytophthora* effector candidates have been found to be prevalently localized in repeat-rich regions of the genome. Due to the high abundance of active TEs, these regions are highly dynamic which might promote expansion or loss of effector genes, thereby enabling rapid evolutionary changes (Haas 2009). Studies on *Mycosphaerella graminicola* revealed that the genome includes eight dispensable chromosomes, the so-called dispensome, which is highly dynamic in field populations and progeny isolates of laboratory crosses (Wittenberg *et al.*, 2009; Goodwin *et al.*, 2011). The chromosomes differ from the core genome with regard to gene and repeat content and appear to have originated by ancient horizontal transfer from an unknown donor.
A genome sequence is the basis for comparative as well as functional genomics analyses, e.g.

transcriptome, proteome and metabolome profiling. It also facilitates marker development for genetic mapping, cloning strategies for genes of interest or generation of mutant lines. In addition, large scale bioinformatics screens to identify candidate genes or sequences of interest become feasible. The fully sequenced genome (Cuomo *et al.*, 2007) and an efficient transformation system enabled reverse and forward genetic screens for pathogenicity mutants of the hemi-biotroph *Fusarium graminearum*, which led to the identification of several genes involved in toxin production (daf genes), fungal development and virulence (reviewed in Kazan *et al.*, 2012). Genome-scale microarrays allow to investigate the *F. graminearum* transcriptome with the aim to better understand the fungal infection process, and proteome studies have already identified a number of secreted proteins that may act as effectors. By screening the genome sequence of *Ustilago maydis*, Kämper *et al.* (2006) discovered gene clusters coding for novel secreted proteins that are crucial for infection and show strong transcriptional activation during biotrophic growth. The best studied example is Pep1, an effector that is essential for biotrophic development (Doehlemann *et al.*, 2009). In addition, the chorismate mutase *Cmu1* was found to be involved in metabolic priming of the host cells during infection (Djamei *et al.*, 2011).

1.6 Powdery mildew pathogenomics: state of the art

Due to the biotrophic lifestyle *Blumeria graminis* can not be cultivated on artificial media, and to date there is no transformation protocol available for *Blumeria graminis*. Thus, the *B.g. hordei* genome sequence, which became public in the course of this work in 2010 (Spanu *et al.*), has enormously stimulated powdery mildew research. It was the first *Blumeria graminis forma specialis* sequenced and was published together with fragmented assemblies of *Erysiphe pisi* (pathogenic on pea (*Pisum sativum*)) and *Golovinomyces orontii* (pathogenic on *Arabidopsis thaliana*). The analysis of the genome sequence revealed new insights such as a large genome size, a very high repetitive content and, reflecting the biotrophic lifestyle, extensive loss of genes which encode enzymes of primary and secondary metabolism (e.g. toxins), carbohydrate-active enzymes and transporters. The most interesting finding was the discovery of the CSEP genes, of which the vast majority is confined to *Blumeria graminis* (compared to *E. pisi* and *G. orontii*) and preferentially expressed in haustoria (79%).

CESPs are currently in the focus of many powdery mildew research projects which aim at characterization of CSEP and their function using proteomics or functional genomics tools such as HIGS (Host induced gene silencing, Nowara *et al.* (2010)). Recently, CSEP0055 was indirectly shown to contribute to fungal infection success, as infection rates in HIGS experiments were significantly reduced when the gene was silenced by an RNAi construct (Zhang *et al.*, 2012). In yeast two-hybrid experiments, CSEP0055 interacted with two barley resistance genes (PR17 and PR1) which suggests that its function as effector is most likely complex.

1.7 Next generation sequencing technologies and their impact on research in biology

Sanger-sequencing, a method for DNA sequencing published in 1977 by Fred Sanger (Sanger *et al.*, 1977), has revolutionised biological research and dominated the sequencing market for almost 30 years. The principle of "sequencing by chain termination" coupled with gel electrophoresis-based size separation dramatically improved earlier DNA sequencing techniques. After 10 more years of technical advancement driven by the vision to sequence the human genome, e.g. capillary gel electrophoresis for fragment separation and fluorescently labelled ddNTPs, the first fully automated capillary sequencer ABI370 appeared on the market (Figure 1.4a). The sequencing of the human genome (Human Genome Project, Consortium (2004)) was done in factory-like sequencing centers with hundreds of capillary sequencers and was completed in 2003. It took 13 years and cost were estimated at 300 million US$.
The enormous cost of the human genome project made clear that cheaper sequencing technologies with higher throughput than the capillary sequencer were needed. In 2005, 454 Life Science published the "sequencing by synthesis" method that uses pyro-sequencing for readout which reduces sequencing reaction volume and multiplies sequencing reactions. Shortly after, competitors came up with a method that provided tenfold more output for a cheaper price but with a shorter read length.

"Next-generation sequencing" (NGS) is a general term for new sequencing technologies that emerged after 2004. The most popular NGS platforms today are Roche 454, Illumina and SOLiD, Illumina dominating the market since years. The platforms differ in their methods of template preparation, sequencing and imaging (reviewed in Metzker (2010)), but they all use a "wash-and-scan" principle: identical strands of DNA anchored to a surface are sequentially flooded with labelled nucleotides, labelled nucleotides are incorporated in the DNA strand, the incorporation reaction is stopped, excess nucleotides are washed away and the incorporated bases are identified by scanning (Figure 1.4b). The individual methodology for template preparation, sequencing and imaging leads to a platform specific sequencing errors mainly caused by biased PCR amplification and dephasing[2]: Roche 454 for example has a problem with homopolymers larger than 5 nucleotides while Illumina is error-prone to GCC sequences.
Most important for the user, however, are the differences in throughput, read length and output format. For example, GS FLX Titanium (Roche 454) produces reads with average lengths of 700bp and a typical throughput of 700 Mb in a runtime of 23 hours[3] whereas the latest HiSeq system from Illumina produces 1.2 billion reads of 2x150bp in 27 hours (=120 Gb)[4]. The output format is crucial for downstream processing and data analysis. SOLiD sequences for example are provided in color-space format which can not be handled by many programs and therefore requires special software.
Short read lengths (Illumina read length is still 150 bp max.) make informatics processes such as assembly and alignments challenging. This issue can be addressed by increased sequencing coverage (resequencing 8-12x, *de novo* 25-70x) and by the use of paired-end or matepair sequences

[2] loss of synchronicity due to the large number of scanning and washing cycles
[3] `454.com/products/gs-flx-system/index.asp`
[4] `www.illumina.com/systems/hiseq\_comparison.ilmn`

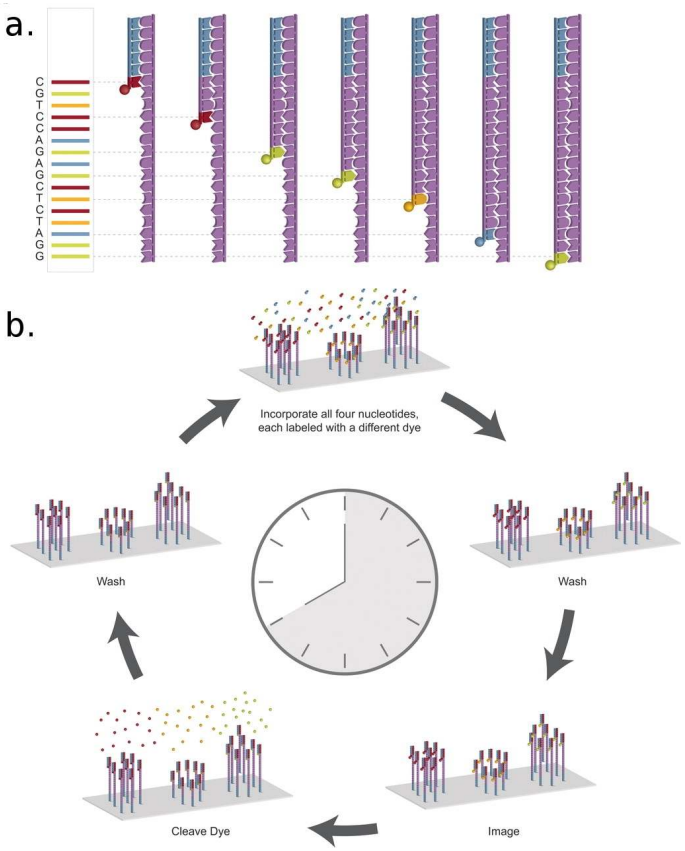

Figure 1.4. Working principle of first- and second generation sequencing technology. **a.** A modern implementation of Sanger sequencing is shown to illustrate differential labeling and use of terminator chemistry followed by size separation to resolve the sequence. **b.** The Illumina sequencing process is shown to illustrate the wash-and-scan paradigm common to second-generation DNA-sequencing technologies. Source: Schadt *et al.* (2010)

(two sequences with a known distance to each other) in addition or instead of single reads. Paired-end/mate-pair sequences are standard for many applications today and are especially helpful when working with repetitive genomes.

Today, the era of "next-next-generation" or "third-generation" sequencing (TGS) has already begun. New technologies aim at providing longer read lengths with high accuracy and reduced PCR amplification and dephasing bias (Schadt *et al.*, 2010) at a low cost (comparison between first-, second- and third generation sequencing is shown in Table 1.2). IonTorrent sequencers for example are small pH meters that detect changes in pH of the sequencing solution caused by the release of a hydrogen ion during the nucleotide incorporation[5]. This principle simplifies the sequencing process substantially as no imaging system is needed, other than that it is still a "wash-and-scan" method that requires template amplification by PCR. Pacific Biosciences (PacBio) sequencing technology takes advantage of the highly efficient and accurate process of DNA replication by observing DNA synthesis in real time[6]. Single molecules of DNA (in contrast to PCR amplified libraries) are used as templates for sequencing. Single DNA polymerases molecules, which are anchored to the bottom of tiny tube-like reaction cells called ZMW (zero-mode waveguide), incorporate labeled nucleotides into the growing DNA strand. Laser light is applied from below, and because it decays exponentially as it enters the cells, it is possible to detect and distinguish single incorporation events against the background of fluorescently labeled nucleotides which diffuse freely though the ZMWs.
At the moment, many TGS technologies underperform regarding raw read accuracy, throughput and cost efficiency. It therefore still remains to be seen whether they will become significantly advantageous over today's well established NGS platforms such as Illumina.

1.8 Bioinformatic challenges created by NGS

NGS technologies offer a wide range of application and have changed the way biologists approach research questions in general: RNA seq for example allows to assemble the transcriptome of an organism even without having a reference sequence available, and characterization of transcripts by sequencing rather than through hybridization to a chip is advantageous especially for large genomes (e.g. plant genomes). However, NGS and its applications also pose challenges for researchers.

The claim that NGS technologies lower sequencing costs does not consider the increasing costs and efforts in bioinformatics. In principle, a sequencing experiment consists of four elements: sample collection, sequencing, data reduction and downstream analysis (Sboner *et al.*, 2011). Comparing the sequencing projects from the pre-NGS (approx. year 2000) to those of today, major costs have shifted away from sequencing to experimental design, data reduction and downstream analysis: Experimental designs using sophisticated sequencing methods (e.g. bisulfite sequencing or methylated DNA immunoprecipitation (MeDIP-Seq)) may require complex molecular and cellular experiments to produce the library for sequencing. Raw sequencing data is extremely big and has to be converted in compressed file structures which are more handy

[5] `http://www.iontorrent.com`
[6] `www.pacificbiosciences.com/products/smrt-technology`

Table 1.2. Comparison of first-, second- and third generation sequencing. Source: Schadt *et al.* (2010)

	First generation	Second generation	Third generation
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS[a] or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800-1000bp)	Short, generally much shorter than Sanger sequencing	Long, 1000bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base, low cost per run	Low cost per base, high cost per run	Low-to-moderate cost per base, low cost per run
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on the technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges

[a] sequencing by synthesis

to work with. This can be the mapping file of the reads to the reference (e.g. Binary Sequence Alignment/Map (BAM)) or a *de novo* assembly of the reads. Further processing is required to obtain "high-level summaries" of the experimental data (Sboner *et al.*, 2011): files which comprise all the information needed for downstream analysis, e.g. a table with all positions of SNPs and InDels in a genome for a variant detection experiment, expression values for all genes of interest for an RNAseq experiment or information about binding sites for a ChIP-seq experiment. The output of the initial data analysis has to be further processed, evaluated and interpreted according to the research question. For example, a set of differentially expressed genes of an RNAseq experiment has to be statistically filtered for those which are significantly differentially expressed. In many cases, profound knowledge in bioinformatics, statistics, genetics and biology

is required to comprehend the complexity of the data.
It is important to note that for many steps of data reduction and downstream analysis no streamlined approaches are available. Therefore, each sequencing project requires considerable efforts to find, install, configure and run appropriate software and computational pipelines. In addition, suitable solutions for data storage and tracking have to be developed and maintained in the lab.

The market for sequencing technology is very dynamic and many of the technologies are not around for long (e.g. SOLiD). This creates problems at the bioinformatics front: The diversity of NGS data characteristics (e.g. read length) and format (sff, fastq, colourspace) demands for a wide range of software such as aligners or assemblers to process the data. Researchers who start to use a new sequencing technology in the early phase of the release to the market often encounter the problem that there is no or very little software for data analysis available. Software development, especially open source software, often lags behind the technical innovation, as developers can only start their work once the technology is known and sequence data are available. If software is available, considerable effort is usually required not only to install the programs and learn how to run them, but also to convert raw data into the form that is accepted by the program. This is usually done via handmade scripts. It is therefore important to consider not only throughput, read length and accuracy when choosing the flavour of NGS for an experiment, but also availability of software for downstream analysis.

The fast upcoming of new NGS technologies resulted in a lack of community standards for data analysis, especially from a statistics perspective. An example is RNAseq used for differential gene expression: For microarrays, there is clear consensus about how to do differential gene expression analysis as technology has been around for several years. For RNAseq however, as the technology is relatively recent, the community is still debating about statistical methods and standards.
Quality control for large scale data is challenging and often based on statistics which is sometimes difficult to comprehend for non-statisticians (e.g. differential gene expression). NGS technologies are nowadays widely used for all kind of analysis and sequencing cost are so low that they become affordable for small labs. Therefore, collaborations between biologists, bioinformatics and statisticians are becoming more important, and guidelines for statistical methodologies are an urgent need.

1.9 Aims of the thesis

A project with the goal to identify fungal interactors of *Pm3* resistance alleles was initiated in 2007. The work presented in this thesis aimed at supporting this intent by producing a high quality genome sequence for *B.g. tritici* isolate 96224. This isolate is one of the parents of a cross which was used to map *AvrPm3f*, the avirulence gene corresponding to *Pm3f*, using a map-based cloning approach.
To asses the extensive repetitive content of the *B.g. tritici* genome, a comprehensive library of the most abundant *B.g. tritici* transposable elements was generated using a low coverage sequencing

run on *B.g. tritici* genomic DNA (described in chapter 2). Preliminary studies on *B.g. tritici* BAC sequences were then undertaken to get a first insight into the characteristics of the *B.g. tritici* genome. The evolutionary relationship between *B.g. tritici* and *B.g. hordei* and the level of conservation on the molecular level was assessed by a comparative analysis of two orthologous loci in these genomes (described in chapter 3).
The main goal of this work was the generation of a reference sequence for *B.g. tritici* isolate 96224 (described in chapter 4) and the analysis of the information encoded in the genome in relation to *B.g. tritici*'s biology as a pathogen. This included the assembly of NGS data and its anchoring to BAC library contigs as well as gene annotation based on models predicted based on homology to *B.g. hordei* genes and *ab initio*. A comparative genomics approach was used to study similarity and differences between *B.g. tritici* and *B.g. hordei* and led to the identification of a set of putative effector genes which might play a role in disease development. Re-sequencing of three additional *B.g. tritici* isolates was undertaken in order to investigate the genetic diversity between isolates of the same *forma specialis* with a special focus on the geographical origin.

CHAPTER 2

A major invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome

Francis Parlange[1,a], Simone Oberhaensli[1,a], James Breen[1], Matthias Platzer[2], Stefan Taudien[2], Hana Šimková[3], Thomas Wicker[1], Jaroslav Doležel[3] and Beat Keller[1]

[1]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[2]Leibniz Institute for Age Research, Fritz Lipman Institute, Beutenbergstrasse 11, 007745 Jena, Germany

[3]Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Sokolovská 6, 77200 Olomouc, Czech Republic

[a]Both authors contributed equally to this work

Published in Functional and Integrative Genomics, Issue 4, page 671-677, December 2011

Supplementary figures and tables can be found in Appendix A.

Abstract

Powdery mildew of wheat (*Triticum aestivum* L.) is caused by the ascomycete fungus *Blumeria graminis* f.sp. *tritici*. Genomic approaches open new ways to study the biology of this obligate biotrophic pathogen. We started the analysis of the *B.g. tritici* genome with the low-pass sequencing of its genome using the 454 technology and the construction of the first genomic bacterial artificial chromosome (BAC) library for this fungus. High-coverage contigs were assembled with the 454 reads. They allowed the characterization of 56 transposable elements and the establishment of the *Blumeria* repeat database. The BAC library contains 12,288 clones with an average insert size of 115 kb, which represents a maximum of 7.5-fold genome coverage. Sequencing of the BAC ends generated 12.6 Mb of random sequence representative of the genome. Analysis of BAC-end sequences revealed a massive invasion of transposable elements accounting for at least 85% of the genome. This explains the unusually large size of this genome which we estimate to be at least 174 Mb, based on a large-scale physical map constructed through the fingerprinting of the BAC library. Our study represents a crucial step in the perspective of the determination and study of the whole *B.g. tritici* genome sequence.

2.1 Introduction

Powdery mildew fungi are some of the most damaging plant pathogens. They affect a wide range of dicotyledonous and monocotyledonous host species and cause significant economic losses in crop plants worldwide (Glawe, 2008). Powdery mildews belong to the family Erysiphaceae in the order Erysiphales (Ascomycota) (Inuma *et al.*, 2007). Their interactions with the host are characterized by the establishment of structures called haustoria inside epidermal plant cells, allowing the pathogen to maintain a parasitic relationship and to take up nutrients from the host. This results in a complete dependence of powdery mildew growth on living plant cells (Glawe, 2008).

The fungal pathogen Blumeria graminis is an ascomycete species subdivided in seven *formae speciales* (ff. spp.), each highly specialized for different host species (Inuma *et al.*, 2007).*Blumeria graminis* f.sp. *tritici* (hereafter called *B.g. tritici*) is the causal agent of powdery mildew on wheat (*Triticum aestivum* L.). Little is known about the biology of this fungus and, therefore, methods and resources are needed to identify genes promoting virulence and determining *B.g. tritici*-wheat interaction and to understand the mechanisms underlying host specialization of *B. graminis*.

Recently, the genome sequencing of *B. graminis* f.sp. *hordei* (*B.g. hordei*), closely related to *B.g. tritici* and causal agent of the powdery mildew of barley (*Hordeum vulgare*), has been completed (Spanu *et al.*, 2010). This work, together with the reports of other obligate biotroph genome sequences (Baxter *et al.*, 2010; Duplessis *et al.*, 2011) revealed genomic hallmarks possibly driven by adaptations to the obligate biotrophic lifestyle. Those include a massive proliferation of transposable elements correlated with expansion of the genome size and the loss of genes which are not essential for the biotrophic lifestyle, such as genes encoding enzymes devoted to plant cell wall degradation or nitrate and sulfur assimilation pathways (Spanu *et al.*, 2010; Baxter *et al.*,

2010; Duplessis *et al.*, 2011).

In order to determine its genomic features, we initiated the exploration of the *B.g. tritici* genome with the construction and characterization of the first bacterial artificial chromosome (BAC) library from this fungus. We fingerprinted the library and produced a physical map of the genome which allowed a first estimation of the genome size. Based on low-pass 454 sequencing of the genome and 20,001 BES representing approximately 7% of the nuclear genome, we were able to build a Blumeria repeat database and to obtain a first insight into the *B.g. tritici* genome.

2.2 Materials and methods

Plant and fungal material

The construction of the BAC library and 454 sequencing were performed using DNA from *B.g. tritici* isolate 96224 (Brunner *et al.*, 2010). Cultures of 96224 were propagated by infecting fresh leaf segments of the susceptible bread wheat cultivar Kanzler, kept on agar supplemented with benzimidazole at a concentration of 30 mg/L.

BAC library

Construction and characterization of the BAC library are described in the supplementary text.

Assembly of a physical contig map of *B.g. tritici*

Fingerprinting was performed at the Instituto di Genomica Applicata (`www.appliedgenomics.org`). High information content fingerprints (HICF) were produced and processed through FPB software (Scalabrin *et al.*, 2009) for fingerprint background removal and GenoProfiler software (You *et al.*, 2007) for removal of contaminants and batch processing of fingerprints into size files that can be input into FPC (Soderlund *et al.*, 1997). Fingerprinted clones were initially assembled using FPC at a Sulston cutoff score of 1e-60 (initial incremental contig build) and Q-clones were split using three DQ steps at slightly lower Sulston scores. Singleton clones were then added to contigs, and ends were merged (when applicable) by increasing the cutoff score by 1e-5 in a stepwise manner to 1e-20 (final cutoff). The approach to control experimentally the accuracy of the FPC assembly is described in supplementary text.

BAC-end sequencing

BAC-end sequencing was made at the Arizona Genomics Institute, University of Arizona (`www.genome.arizona.edu`). Sequencing was performed at both ends. Sequence trimming was conducted by processing trace files using the Phred program for base calling and a quality score of 20 (Ewing *et al.*, 1998). Vector sequences were masked using CROSS_MATCH (`www.genome.washington.edu`) and removed from the analysis. Only reads with a length of at least 100 bp were retained, providing 20,001 high-quality BAC-end sequences.

Construction of the *Blumeria* repeat database

The low-pass genome sequencing of the *B.g. tritici* isolate 96224 was performed using the GS FLX platform (Roche) (Supplementary Text). Reads were assembled using the MIRA software with default settings for assembly of 454 sequences. Contigs with a 10-25x coverage and a minimal length of 7 kb were used for the manual characterization of full-length transposable element (TE) sequences. The strategy for the identification of TEs was the following: BLASTN

and BLASTX searches (Altschul *et al.*, 1997) against specialized databases such as RepBase (`www.girinst.org`) and TREP (`wheat.pw.usda.gov/ITMI/Repeats/`) were performed in order to reveal typical features characterizing the different superfamilies of TEs. Long interspersed nuclear elements (LINE) were identified by their generally well-conserved ORF2 sequence. The presence of associated ORF1 and poly-A sequences allowed further identification of complete elements. Short interspersed nuclear elements (SINE) were identified by the presence of internal A and B promoter boxes necessary for RNA polymerase III binding as well as a poly-A tail at the 3′ end. For long terminal repeat (LTR) retrotransposons, typical patterns of the terminal repeats were revealed using DOTTER (Sonnhammer and Durbin, 1995). Target site duplications and LTR borders were determined manually. The classification into copia or gypsy superfamilies was done according to similarity of the ORF-encoded proteins with the PTREP database, and their internal organization within the element (Wicker *et al.*, 2007). Additionally, we used contigs of the *B.g. hordei* draft genome (version June 2007) which were made available for us by the BluGen consortium (`www.blugen.org`) for homology search to identify the *B.g. hordei* homologs of *B.g. tritici* repeats. In order to reduce redundancy within the different families, we set a threshold of 80% similarity at the nucleotide level for the definition of a family. Finally, elements were named according to the nomenclature of Wicker *et al.* (2007).

BES analysis

The 20,001 BES were first analyzed for their repeat content through BLASTN and BLASTX searches (Altschul *et al.*, 1997) against the *Blumeria* repeat database. Only hits with a minimal alignment of 100 bp, 80% of nucleotide identity (for BLASTN) and an E value <10 e-10 (for BLASTX) were considered. For the identification of additional high-copy sequences, sequences matching the repeat database were removed, and the remaining ones were searched against themselves using the same BLASTN parameters.

Access to sequence data

All BAC-end sequences can be accessed through accession numbers FR776010 to FR796010 in the EMBL nucleotide sequence database. An FTP server (address available on request) provides access to the complete set of sequences of the 56 identified *Blumeria* repeats (files Bg_repeats_fasta and Bg_repeats_hypothetical_proteins_fasta).

2.3 Results

Fingerprinting of the *B.g. tritici* BAC library provides a physical map of the genome and an estimate of the minimal genome size

A large insert BAC library was constructed with *B.g. tritici* reference isolate 96224 (Supplementary Text). Fingerprinting of the complete library (12,288 clones) generated 6,831 HICF which were assembled to produce 266 BAC contigs (Table 2.1). Only 146 (2.1%) BAC clones remained as singletons. The largest contig is 5.8 Mb, and 50% of the assembly is contained in contigs larger than 1 Mb. By comparison with experimentally tested overlaps of BAC clones at two genomic regions (Supplementary Figure A.3 and Supplementary Table A.1), we could confirm the accuracy of the fingerprint assembly and its relevance for establishing contigs spanning large genomic regions. The total length of the assembly is 174 Mb, giving a first estimate of the *B.g. tritici* minimal genome size.

Table 2.1. Characteristics of the *B.g. tritici* contig assembly

Total clones	12,288
Useful fingerprints	6,831
Assembled contigs	266
Clones in contigs	6,685
Singletons	146
Maximum nr of clones per contig	325
Largest contig	5,825 kb
N50 (nr contigs)	51
Length of N50 contig	1,002 kb
Total length of assembly	174 Mb

Construction of a *Blumeria* repeat database

In order to study the fraction of repetitive DNA in the *B.g. tritici* genome, we established a Blumeria repeat database, exploiting two datasets of sequence information. First, whole genome sequencing of the *B.g. tritici* genome was carried out by one full 454 GS FLX run. This resulted in 491,163 reads with an average size of 226 bp. Assembly of these reads produced 39,363 contigs and contigs with a very high coverage were selected, as this indicates sequences corresponding to high-copy repeats. Additionally, we also exploited few contigs belonging to the first *B.g. hordei* draft genome sequence (version June 2007) which were made available to us by the BluGen consortium (`www.blugen.org`).
Composition of the Blumeria repeat database is presented in Table 2.2. We identified 20 families

of LINEs and two *B.g. tritici* SINEs, Bgt_RSX_Yhi and Bgt_RSX_Lie, homologs of the previously characterized *B.g. hordei* SINE elements EGH-24-1 (Rasmussen *et al.*, 1993) and EG-R1 (Wei *et al.*, 1996), respectively. A total of 27 LTR retrotransposons were found (Table 2.2), of which 13 families could be classified as members of the gypsy superfamily and nine as members of the copia superfamily. Five sequences showed characteristics of solo LTRs, but the complete retrotransposon they originated from could not be characterized. Finally, seven sequences exhibited characteristics of TE and a high-copy number, but could not be classified into any order of repeat ("unclassified" in Table 2.2). Among them were two *B.g. tritici* sequences for which we could identify two homologous sequences in *B.g. hordei* (both *B.g. tritici* and *B.g. hordei* homologs are in the database).
In conclusion, our *Blumeria* repeat database is composed of 56 TE families, including some elements which are conserved in *B.g. tritici* and *B.g. hordei* (Table 2.2).

Table 2.2. Transposable element families of the *Blumeria* repeat database and representation of the superfamilies in the BES dataset

Order	Superfamily	Families in the database	Percentage of the BES databse in length
LINE		20	21.6
SINE		2	3.0
LTR retrotransposons	Gypsy	13	8.3
	Copia	9	8.3
	Solo LTRs	5	0.6
Unclassified		7	6.0
Total		56	47.8

BAC-end sequencing and TE content analysis

All the 12,288 BAC clones of the library were sequenced from both ends. After trimming the individual sequencing reads for length (threshold of 100 bp) and low-quality bases, vector and bacterial contaminant sequences were eliminated. In the end, the *B.g. tritici* BAC-end database consisted of 20,001 sequences with an average read length of 633 bp (Supplementary Figure A.4). The total BES length is 12,662,922 bp with an average GC content of 44.3%. This large dataset of representative, random sequence was subsequently used to analyze the composition of the *B.g. tritici* genome.
Sequences corresponding to TEs were first identified in the 20,001 BES by BLASTN search against our *Blumeria* repeat database. The cumulative length of sequences with homology to the 56 repeat families represented 24.1% of the BES database (Supplementary Figure A.5), suggesting that the characterized repeat families could contribute approximately one fourth of the genome.

The ten most abundant elements represented half of the TE fraction (49.8%), and accounted for around 12% of the genome (Supplementary Figure A.5). Five LINE elements represented all together 6.2% of the genome. The most abundant element of all was the SINE Bgt_RSX_Yhi (2%).

We then masked the sequences matching the *Blumeria* repeat database at the nucleotide level, and performed with the remaining sequences a second search against the *Blumeria* repeat database at the protein level, in order to evaluate the representation of TE superfamilies. A cumulative length representing 23.7% of the BES set gave hits. Taken together with the previous analysis, the fraction of the BES set matching TEs of the *Blumeria* repeat database is 47.8%, i.e. 6.04 Mb. The analysis of these sequences revealed the predominance of non-LTR retrotransposons over LTR retrotransposons, mainly due to LINE elements (Table 2.2).

In order to identify additional unknown repeats, we masked all the sequences which previously matched our repeat database at the nucleotide and protein level, and kept only the BES if the remaining unmasked sequence was longer than 50 bp. This resulted in 13,270 remaining BES which were searched against themselves by BLASTN. Repeats or high-copy sequences were defined as sequences with at least two copies in the 13,270 BES set. Considering that the complete BES database represents 7.2% of the *B.g. tritici* minimal genome size, a high-copy sequence according to our definition would then be expected to occur in more than 28 copies in the genome. This search revealed 8,880 high-copy BES with a total length of 4.74 Mb. Together with the 6.04 Mb matching the *Blumeria* repeat database, we estimate the total repeat content in the BES database, and by extension in the *B.g. tritici* genome, to be 85%.

2.4 Discussion

In this paper, we report on the construction and characterization of the first *B. graminis* f.sp. *tritici* large insert BAC library. The majority of BAC libraries constructed from fungal or oomycete pathogens have a relatively small average insert size between 40 and 80 kb, and those constructed from the barley powdery mildew *B.g. hordei* were reported to have average insert sizes of 30 and 41 kb (Ridout and Brown, 1999; Pedersen *et al.*, 2002a). The *B.g. tritici* BAC library consists of 12,288 clones of 115 kb on average with 87% of the inserts larger than 100 kb. This result is remarkable for DNA obtained from a true obligate biotrophic fungus which cannot be cultivated in vitro, and is comparable with the largest libraries reported for ascomycete or oomycete species (Zhu *et al.*, 1997; Zhang *et al.*, 2006; Chang *et al.*, 2007). With a 7.5x coverage of the genome, our BAC library thus represents a powerful tool for the exploration of the *B.g. tritici* genome.
Taking advantage of this library, we could show that *B.g. tritici* possesses an expanded genome of at least 174 Mb, much larger than what is commonly observed for fungal genomes (Gregory *et al.*, 2007). This observation is in accordance with the recently reported genome size of the closely related barley powdery mildew pathogen *B.g. hordei*, which is estimated to be 120 Mb (Spanu *et al.*, 2010), and demonstrates that the *formae speciales* of the *B. graminis* species have an atypically large genome size. The high percentage of repeats in *B.g. tritici* (85%) seems to be the explanation for the unusually large size of its genome, which is possibly also true for the genome of *B.g. hordei* as hypothesized by Spanu *et al.* (2010). We observed that non-LTR retrotransposons in the form of LINEs are predominant over LTR retrotransposons in the *B.g. tritici* genome. SINEs are also surprisingly abundant in *B.g. tritici* and could represent at least 3% of the genome, although they are relatively small in size (Wicker et al. 2007). Similarly, Spanu *et al.* (2010) observed that LINEs and SINEs are largely predominant over LTR retrotransposons. This picture is different than what was recently reported in other repeat-rich oomycete and fungal genomes such as *Hyaloperonospora arabidopsis* (Baxter *et al.*, 2010), *Melampsora larici-populina* and *P. graminis* f.sp. *tritici* (Duplessis *et al.*, 2011). In *B.g. hordei* as well as in *H. arabidopsis*, only a small fraction of class II transposable elements was detected (Spanu *et al.*, 2010; Baxter *et al.*, 2010), which is not the case for *M. larici-populina* and *P. graminis* f.sp. *tritici* where the proportion of class I and class II elements is more equal (Duplessis *et al.*, 2011).

The very stringent parameters we used to assess the fraction of repeat DNA (80% identity) indicates that repeat copies are very similar, which could suggest that proliferation of repetitive DNA in *B.g. tritici* is the consequence of a high rate of recent transposon activity. Recently, Oberhaensli *et al.* (2011) sequenced and annotated three *B.g. tritici* BAC clones. They found a large difference of TE content in a comparative analysis with *B.g. hordei*, indicating that indeed most of the TE activity in the two genomes occurred after divergence of the two *formae speciales*, around 10,000,000 years ago. In the same study, it was found that TEs accounted for 48.8% and 51.4% of the contigs length, respectively. However, those clones were specifically screened to encompass gene-containing regions. On a third locus, TEs were shown to occupy up to 69% of the sequence (F. Parlange, unpublished results), which is closer to the estimation presented in the current study. This suggests that repeated elements may not be equally distributed along the genome, and proves the importance of generating large and randomly dispersed sets of sequences to draw an accurate picture of the composition of large and highly repetitive genomes.

The reports on genome sequences from three powdery mildew species, including *B.g. hordei*, *Erysiphe pisi*, and *Golovinomyces orontii* (Spanu *et al.*, 2010), and the "downy mildew" *H. arabidopsis* (Baxter *et al.*, 2010) highlighted striking signatures of convergent evolution to an obligate biotrophic lifestyle, in particular marked by an unusually expanded genome size correlated with a proliferation of transposable elements. Recently, the same observation was reported in two other obligate biotrophic parasites, the rust fungi *M. larici-populina* and *P. graminis* f.sp. *tritici* (Duplessis *et al.*, 2011). Those observations in different evolutionary lineages support the hypothesis of Spanu *et al.* (2010) that large genome size and high repetitive DNA content are common hallmarks associated with obligate biotrophy. Transposable elements affect the genome by their ability to move and replicate. They can generate high levels of genetic variation independent of sexual recombination, and could contribute to genome flexibility responsible for rapid adaptation of populations to selection imposed by resistance genes in the case of phytopathogenic fungi or to environmental constraints for symbionts. The genomes of the basidiomycete fungus *Laccaria bicolor* and the ascomycete *Tuber melanosporum*, which form ectomycorrhizal symbiosis with their host plant, were also reported to be 65 and 125 Mb respectively, with a high proportion of repeats (21% and 58% respectively; (Martin *et al.*, 2008, 2010a).

A convergent biotrophic adaptation was also observed at the genetic level, with a common reduction of genes which are not essential for the biotrophic lifestyle, such as genes encoding enzymes involved in the primary and secondary metabolism (Spanu *et al.*, 2010), enzymes devoted to plant cell wall degradation (Spanu *et al.*, 2010; Baxter *et al.*, 2010; Duplessis *et al.*, 2011) and transporters (Spanu *et al.*, 2010). The absence of genes involved in the inorganic nitrate and sulfur assimilation pathways also seems to be a feature of obligate biotrophic genomes (Spanu *et al.*, 2010; Baxter *et al.*, 2010; Duplessis *et al.*, 2011). However, little is still known about the molecular mechanisms involved in the establishment of the interaction between obligate biotrophic fungi and their hosts. Investigations on those aspects represent the major challenge in the study of this class of pathogens.

The future sequencing and annotation of the complete *B.g. tritici* genome are the next steps in the exploration of this genome. Sequencing can now be considered through next generation sequencing technologies (Nowrousian *et al.*, 2010), and the physical map and BES generated in this study should greatly facilitate assembly of the genome. The updated *Blumeria* repeat database will also help to overcome difficulties related to the massive presence of TEs and simplify the identification of gene coding sequences. This should provide the opportunity for comparative studies with the other recently sequenced powdery mildew genomes or, at a broader scale, with obligate biotrophic genomes, and contribute to the understanding of the molecular features determining the pathogenesis of those parasites.

Acknowledgments

We would like to thank Gabriele Büsing for the excellent technical assistance. We thank the Blugen consortium (`www.blugen.org`) and especially Dr. P. Spanu for access to the barley powdery mildew genome sequences. This work was supported by the Swiss National Science Foundation grant 3100A-127061/1 (BK), an advanced grant of the European Research Council (durable

resistance 249996, BK) and by European Union grant no. ED0007/01/01 Centre of the Region Haná for Biotechnological and Agricultural Research (HŠ, JD).

2.5 Supplementary text

2.5.1 DNA isolation

For 454 sequencing, conidiospores were ground with glass beads (1.7-2.0 mm) in a Mixer Mill MM300 (Retsch GmbH), then mixed with 2 ml of pre-warmed (65°C) 2x CTAB buffer (2% CTAB, 200 mM Tris/HCl pH 8.0, 20 mM EDTA, 1.4 M NaCl, 1% PVP, 0.28 M β-Mercaptoethanol) and incubated for 1h at 65°C. The volume was adjusted to 6 ml with 2x CTAB. The homogenate was extracted with an equal volume of dichloromethane : isoamylalcohol (24:1) and centrifuged for 15 min at 2,800 rpm. This step was repeated twice. RNA was digested by RNase A (10 mg/Âţl). DNA was precipitated with 0.7 volume of cold isopropanol and centrifuged for 10 min at 3,200 rpm. The pellet was washed for 15 min with Solution I (76% ethanol, 200 mM sodium acetate, 100 mM Tris/HCl pH 7.4), then 2 min with Solution II (76% ethanol, 10 mM NH4 acetate) and centrifuged for 2 min at 2,800 rpm. DNA was air-dried and resuspended in 50 μl TE (10 mM Tris, 1 mM EDTA) buffer.
For the BAC library construction, High Molecular Weight (HMW) DNA was prepared according to the protocol used by (Pedersen *et al.*, 2002b) with some modifications. One gram of conidiospores was lyophilized (220 mg dried material), washed twice in 50 mM EDTA (pH 8.0), 0.5% Tween 20 followed by centrifugation for 10 min at 3,500 rpm. A third wash was performed without Tween 20. The pellet was resuspended in 100 μl of 50 mM EDTA (pH 8.0) containing a cocktail of lysing enzymes (Sigma L1412) at 48 mg/ml. The suspension was incubated at 40°C during 20 min, mixed with an equal volume of pre-warmed 1.8% Incert Agarose (Lonza, Rockland, USA) prepared in 50 mM EDTA (pH 8.0) and transferred to plastic moulds at 4°C. After solidification, agarose plugs were incubated at 37°C for 20h in LET solution [0.5 M EDTA (pH 8.0), 10 mM Tris-HCl (pH 8.0), 5 mM DTT] containing 48 mg/ml of lysing enzymes. Plugs were then incubated 2 x 24h at 50°C in NDA solution [0.5 M EDTA, 10 mM Tris-HCl (pH 9.5), 1% sodium N-lauroyl sarcosinate] with 1mg/ml proteinase K. Plugs were washed 3 x 1 h in 100 mM EDTA (pH 8.0). For long time storage, plugs were equilibrated in 70% ethanol during 8h at room temperature and stored at -20°C.

2.5.2 BAC library construction and characterization

Agarose plugs were equilibrated in ice-cold TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) buffer (6 ml/plug) for 20h to remove ethanol. Before digestion, plugs were washed 3 x 1h in ice-cold TE supplemented with 100 mM PMSF, then 3 times in TE without PMSF, and finally stored overnight in TE at 4°C. The library construction was performed in two rounds, using 5 and 6 plugs for the first and the second round, respectively.
The library was constructed as described in Peterson et al. (2000) with some modifications (Simková *et al.*, 2011). To evaluate digestibility of HMW DNA, preliminary tests were performed on 3 plugs cut into 9 pieces each: the first plug was a control, the second was partially digested by 10 U/ml HindIII during 20 min, and the third plug was completely digested by 100 U/ml HindIII during 6h. For library construction, plugs were cut into pieces, distributed three by three into tubes and partial digestion of the HMW DNA was performed with 4 to 10 U/ml HindIII

(6.5 U/ml on average). For the first 5 plugs, the partially digested DNA was size-separated by PFGE (Pulsed-Field Gel Electrophoresis) in a 1% SeaKem Gold Agarose gel (Lonza, Rockland, USA) in 0.25x TBE under the following conditions: 12.5 °C, 6V/cm, switch time 1-50 s, 17h. The size fraction of 100-150 kb was excised from the gel and subjected to a second step of size selection in a 0.9% SeaKem Gold Agarose gel in 0.25x TBE (12.5 °C, 6V/cm, switch time 3 s, 17h). The size fraction of 90-150 kb was excised from the gel and split into fractions of 100-120 kb (B) and 120-150 kb (M1), respectively. The DNA of particular fractions was electroeluted from the gel and amount of the released DNA was estimated in standard 1% agarose gel by comparing with dilution series of phage λ. Each of the fractions was used to ligate with HindIII-digested cloning-ready pIndigoBAC-5 vector (Epicentre, Madison, USA) in1:3.6 molar ratio (DNA:vector). For the second batch of 6 plugs, only the M fraction (M2) was used for ligation. The recombinant vector was used to transform E. coli ElectroMAX DH10B competent cells (Invitrogen, Carlsbad, USA). Bacterial colonies were picked using Qbot (Genetix, New Milton, UK) and ordered in 32 x 384-well plates filled with 75 μl of freezing medium (2YT supplemented with 6.6% glycerol and 12.5 mg/l chloramphenicol). The BAC library has been stored at -80°C, and is permanently maintained at the Institute of Experimental Botany in Olomouc.
Three hundred BAC clones (60 from the B, 160 from the M1 and 80 from the M2 fraction) were used to estimate the average insert size. The DNA was isolated using standard alkaline lysis method and digested with NotI (0.02 U/μl). DNA fragments were separated in 1% agarose gel in 0.25x TBE buffer by PFGE at 12.5°C, 6V/cm, switch time ramp 1-40 s, 15h. Insert sizes were estimated by comparing with Lambda Ladder PFG Marker and MidRange Marker I (New England Biolabs, Beverly, USA).
For the screening of the library, three dimensional (3-D) pools have been prepared. The 32Âă-plates were subdivided into 4 stacks (8 plates each). Clones of each stack were combined to create a superpool of clones. Further, 48 3-D pools (8 plate, 16 row, 24 column) were prepared for each stack. Thus, the entire library is represented by 192 3D-pools. The pools were processed as described in (Simková *et al.*, 2011).
For fingerprinting and BAC-ends sequencing, two replica of the BAC library were prepared by inoculating new 384-well plates filled with freezing medium with clones of the master copy. After 20h growth at 37°C, the replica were frozen and sent for fingerprinting and BAC-ends sequencing.

2.5.3 Assessment of FPC assembly accuracy

The two loci used to control the accuracy of the FPC assembly correspond to overlapping BAC clones identified by PCR-screening of the 3-D pools (plates 1 to 16, 3.75x genome coverage) using distinct molecular markers. The first region is called locus 2 according to (Oberhaensli *et al.*, 2011) who previously described the screening approach at this locus. For the second region, we exploited a genetic map of Bg tritici (Parlange and Keller, unpublished data) and chose arbitrarily the AFLP marker GTCA_E4 (Forward primer: CAAAGGTAATTTCATCCACTGGT; Reverse primer: CATGACATGAGCAATATCAATACA) for the screening of the 3-D pools. Accordingly, the locus was named GTCA_E4. Supplementary Figure A.3a and b were produced by parsing the FPC files using the WICKERsoft software (available on request).

CHAPTER 3

Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene colinearity, dates divergence and indicates host-pathogen co-evolution

Simone Oberhaensli[1,a], Francis Parlange[1,a], Jan P. Buchmann[1], Fabian H. Jenny[1,b], James C. Abbott[2], Timothy A. Burgis[2], Pietro D. Spanu[2], Beat Keller[1] and Thomas Wicker[1]

[1]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[2]Department of Life Sciences, Imperial College London, Imperial College Road, London SW7 2AZ, UK

[a]Both authors contributed equally to this work

[b]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

Published in Fungal Genetics and Biology, Volume 48, Issue 3, Pages 327-334, March 2011

Supplementary figures and tables can be found in Appendix B.

Abstract

The two fungal pathogens *Blumeria graminis* forma specialis *tritici* (*B.g. tritici*) and *hordei* (*B.g. hordei*) cause powdery mildew specifically in wheat or barley. They have the same life cycle, but their growth is restricted to the respective host. Here, we compared the sequences of two loci in both cereal mildews to determine their divergence time and their relationship with the evolution of their hosts. We sequenced a total of 273.3 kb derived from *B.g. tritici* BAC sequences and compared them with the orthologous regions in the *B.g. hordei* genome. Protein-coding genes were colinear and well conserved. In contrast, the intergenic regions showed very low conservation mostly due to different integration patterns of transposable elements. To estimate the divergence time of *B.g. tritici* and *B.g. hordei*, we used conserved intergenic sequences including orthologous transposable elements. This revealed that *B.g. tritici* and *B.g. hordei* have diverged about 10 million years ago (MYA), two million years after wheat and barley (12 MYA). These data suggest that *B.g. tritici* and *B.g. hordei* have co-evolved with their hosts during most of their evolutionary history after host divergence, possibly after a short phase of host expansion when the same pathogen could still grow on the two diverged hosts.

3.1 Introduction

Powdery mildew fungi are pathogens which belong to the Erysiphales (Ascomycota) and infect a wide range of angiosperm plants. About 650 powdery mildew species are known that occur on almost 10,000 host species (Glawe, 2008). These pathogens are obligate biotrophs: they depend on living plant cells for survival and reproduction. By forming a haustorium that invaginates the epidermal cell of the host plant, the fungus establishes a specific feeding structure that enables the uptake of host nutrients.

Wheat and barley powdery mildew disease is a major problem in the crop producing regions of Asia, northern Europe, north and east Africa as well as in north and south America (Curtis *et al.*, 2002). It has negative effects on yield quality (K. Everts, 2001) and quantity (Conner *et al.*, 2003) and consequently leads to large economic damage. The causal agents, *Blumeria graminis* forma specialis *tritici* (*B.g. tritici*) and *Blumeria graminis* forma specialis *hordei* (*B.g. hordei*), respectively, belong to the cereal powdery mildews (*Blumeria graminis* (DC) Speer), a single species that comprises eight *formae speciales* (ff. spp.) (Inuma *et al.*, 2007). They can be distinguished by their host specialization because they are restricted to a single host. There is a large interest in studying the molecular basis of the powdery mildew-host interaction, since this knowledge could lead to a better understanding of resistance mechanisms. The availability of whole genome information and emergence of next-generation sequencing techniques is facilitating *in silico* approaches to research questions that are difficult to investigate by molecular methods with organisms that are classically considered intractable.

Infection of barley by *B.g. hordei* has been studied intensively during the last 20 years (reviewed in Zhang *et al.* 2005). The mapping and cloning of powdery mildew resistance (Seeholzer *et al.*, 2010) and avirulence genes ((Pedersen *et al.*, 2002a; Ridout *et al.*, 2006)) improved our understanding of the molecular mechanisms of R-gene dependent powdery mildew resistance in bar-

ley (Shen *et al.*, 2007). The *B.g. hordei* genome has been sequenced and annotated in great detail (`www.blugen.org`, (Spanu *et al.*, 2010)). In addition, large scale analysis of the *B.g. hordei* proteome (Noir *et al.*, 2009; Godfrey *et al.*, 2009; Bindschedler *et al.*, 2009) represents valuable resources for studies on haustoria function. We have initiated research on *B.g. tritici* genomics by producing a high quality BAC library from asexual conidia with 8x genome coverage (Parlange *et al.*, 2011). The *B.g. tritici* genome size and the repetitive DNA content (about 70%) is comparable to the genome of *B.g. hordei* (Parlange *et al.*, 2011).

Wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*), the respective hosts of *B.g. tritici* and *B.g. hordei*, diverged about 12 million years ago (MYA) from their last common ancestor (Supplementary Figure B.1), Chalupska *et al.* 2008; SanMiguel *et al.* 2002). They belong to the tribe of the Triticeae along with rye (*Secale cereale*), *Aegilops* and other grass species. The host specificity of *B.g. tritici*, *B.g. hordei* and the other *Blumeria graminis* ff. spp. raises questions about the evolutionary relationship of these pathogens with their hosts. There are conflicting hypotheses concerning their evolution. In one model, *B.g. tritici* and *B.g. hordei* diverged from a common ancestor at the same time as wheat and barley and subsequently co-evolved with their present hosts. An alternative possibility is a "host-jump" of a former non-host pathogen to wheat or barley (Stukenbrock and McDonald, 2008) in relatively recent times, followed by rapid host specialization and a subsequent shift to the closely related cereal later on ("host-shift"; Stukenbrock and McDonald 2008).

In the last decade, several studies have tackled the evolutionary relationship of *Blumeria graminis* ff. spp. Wyand and Brown (2003) compared rDNA-ITS (internal transcribed sequences) and the β-tubulin gene of the wheat, barley, oat and rye powdery mildew pathogens. Because there were discrepancies between phylogenetic trees of four *B. graminis* ff. spp. and their hosts, they considered the co-evolution hypothesis to be unlikely. Instead, they suggested that divergence of B. graminis ff. spp. in agriculture has taken place within the past 14,000 years (Supplementary Figure B.1). In contrast, the study of Takamatsu (2004) resulted in a phylogenetic tree that indicates a *B.g. tritici-B.g. hordei* divergence of roughly 10 MYA. This estimate was based on 600 bp of the 28S rDNA gene. Furthermore, the analysis of (Inuma *et al.*, 2007) suggests that the split of the *Hordeum* and *Triticum* clades of *B. graminis* has happened 4.6 million years ago. This result was based on applying Takamatsu's mutation rate on rDNA (ITS). This wide range of divergence time estimates reflects the challenges of fungal molecular dating in the absence of fossil records (Takamatsu, 2004), reliable mutation rate estimates and sufficient phylogenetically informative sequences (Wyand and Brown, 2003).
Substitution rates can vary significantly among gene-coding sequences, as an effect of selective pressure on specific loci or a particular lifestyle of a species. In contrast, the neutral mutation rate of protein-coding genes (the rate of synonymous substitutions at the third base of the codon) is surprisingly constant across plants, animals, bacteria and fungi (Kasuga *et al.*, 2002). Phylogenetic distances can only be determined by using a molecular clock when the nucleotide substitution rate of the compared sequences is constant (Rutschmann, 2006). Unlike protein-coding genes or sequences with regulatory functions, intergenic regions including pseudogenes and inactive transposable elements (TE) are assumed to be free from selection pressure. Therefore, nucleotide substitutions in these sequences occur at a neutral rate (Petrov, 2001). This concept was the basis for the estimation of long terminal repeat retrotransposon (LTR) insertion

time in plants (SanMiguel *et al.*, 1998) and fungi (Martin *et al.*, 2010a).

Here, we specifically focus on the evolutionary relationship of *B.g. tritici* and *B.g. hordei*. We compared two loci from *B.g. tritici* BAC sequences with the corresponding regions in the *B.g. hordei* genome and calculated their phylogenetic distance based on conserved intergenic sequences or transposable elements. We found colinearity of orthologous genes while intergenic regions were poorly conserved and heavily populated with transposable elements. From our divergence time estimate of about 10 MYA, we conclude that *B.g. tritici* and *B.g. hordei* have co-evolved with their hosts for most of the time after the divergence of wheat and barley.

3.2 Materials and methods

BAC clone selection, sequencing and annotation

A whole genome shotgun sample of *B.g. tritici* isolate 96224 was SOLEXA sequenced (C. Ridout, John Innes Centre UK) and represents a database of short single sequences (35 bp). *B.g. hordei* sequences were provided by the Blugen consortium and originate from the Arachne assembly (version June 2007). Primer pairs (Supplementary Table 2, `www.sciencedirect.com/science/article/pii/S108718451000188X`) were designed based on putative coding sequences present on the *B.g. hordei* scaffolds and then used to screen the *B.g. tritici* BAC library (Parlange *et al.*, 2011) to identify clones containing the orthologous sequences. After screening the 3D-DNA pools, candidate clones were confirmed by PCR and insert size was estimated by digestion with NotI on pulse field gel electrophoresis (PFGE). BAC clones 2p10 (∼95 kb, locus 2), 1f12 (∼130 kb, locus 1) and 12c21 (∼120 kb, locus 1) were shotgun sequenced using Sanger method and assembled using Phred-phrap (available at phrap.org). Conserved regions between the *B.g. tritici* BACs and the *B.g. hordei* scaffolds were determined with DOTTER (Sonnhammer and Durbin, 1995). Transposable element annotation on *B.g. tritici* and *B.g. hordei* sequences was done by BLASTN/BLASTX against a Blumeria repeat database (Parlange *et al.*, 2011) followed by manual verification of hits with DOTTER. Genes were identified by BLASTX against yeast (`www.yeastgenome.org`) and Magnaporthe oryzae protein database (`www.broadinstitute.org/annotation/genome/magnaporthe\_grisea/MultiHome.html`; release 5). Inhouse perl scripts were used to visualize the comparison of the annotated sequences. Sequence data from this article have been submitted to Gen-Bank under accession Nos. HQ437159 and HQ437160.

Calculation of divergence time

Intergenic sequences that were conserved between *B.g. tritici* and *B.g. hordei* and were at least 1 kb up- or downstream of genes were aligned with the program WATER (EMBOSS package, `http://emboss.sourceforge.net`). Kimura 2-parameter criterion (Kimura, 1980) was applied to determine the transition/transversion ratio. Based on this, the divergence time was estimated using a mutation rate of 1.3 x 10^{-8} (Ma and Bennetzen, 2004). All scripts used in this study are available upon request.

3.3 Results

Screening for orthologous loci in *B.g. tritici* and *B.g. hordei* genomes

Since there was no assembled genome sequence available for *B.g. tritici*, we decided to use single clones of a *B.g. tritici* BAC library for a comparative analysis with sequences of the *B.g. hordei* draft assembly. The *B.g. tritici* BAC library was produced from high molecular weight DNA of conidia and is of high quality (8x coverage, 12,288 clones, 115 kb average insert size, Parlange *et al.* (2011)). In order to find gene-rich loci for comparative analysis, we first screened the *B.g. hordei* genome (version June 2007, `www.blugen.org`) for scaffolds with regions larger than 30 kb that contained genes but no sequence gaps. As we were expecting a high content of repetitive elements in the sequences, the candidate scaffolds from *B.g. hordei* were first subjected to a BLASTN search against a dataset of *B.g. tritici* SOLEXA shotgun reads (single reads of 35 bp) that represent about 6x genome coverage. This allowed us to distinguish high- and low-copy regions on the *B.g. hordei* scaffolds without annotating them: regions with hits to many SOLEXA reads were considered repeat-rich, and those with low SOLEXA coverage as potentially gene-containing. We chose two gap-free regions with multiple low-copy regions (supposedly genes) for further analysis. They had a size of 58.6 kb (largest region without gaps) and 67.6 kb, respectively, and they will be referred to as *B.g. hordei* locus 1 and *B.g. hordei* locus 2 hereafter. The two loci were subsequently searched for the presence of genes by a BLASTX search against fungal protein databases. After gene annotation, primers were designed to screen the *B.g. tritici* BAC library. Three *B.g. tritici* BAC clones were identified and their inserts were completely sequenced: clones 1f12 and 12c21 are partially overlapping and correspond to *B.g. hordei* locus 1, and clone 2p10 contains the region orthologous to *B.g. hordei* locus 2 (Figure 3.1).

The *B.g. tritici* BAC sequences and the *B.g. hordei* loci were fully annotated using homology to genes from yeast and M. oryzae (5. release, Dean *et al.* 2005). During the process of transposable element (TE) annotation, we found that *B.g. tritici* and *B.g. hordei* have very similar types of TEs, namely class 1 retrotransposons such as long terminal repeat (LTR) retrotransposons, long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Details on superfamilies and their relative contribution to the total sequences of the analyzed loci are given in Table 3.1. Individual TE families from *B.g. tritici* and *B.g. hordei* were up to 90% identical at the DNA level. Therefore, TEs in the *B.g. tritici* BAC sequences and in the *B.g. hordei* scaffolds were annotated using a *B.g. tritici* repeat database which mainly includes TEs that were detected on high-coverage scaffolds of a preliminary assembly of *B.g. tritici* whole genome shotgun 454 reads (Parlange *et al.*, 2011).

Locus 1 shows large size differences of intergenic sequences

Locus 1 corresponds to the overlapping *B.g. tritici* BACs 1f12 and 12c21 and to *B.g. hordei* scaffold sc3 (Figure 3.1a). The total length covered by BACs 1f12 and 12c21 is 218.4 kb. Annotation of this

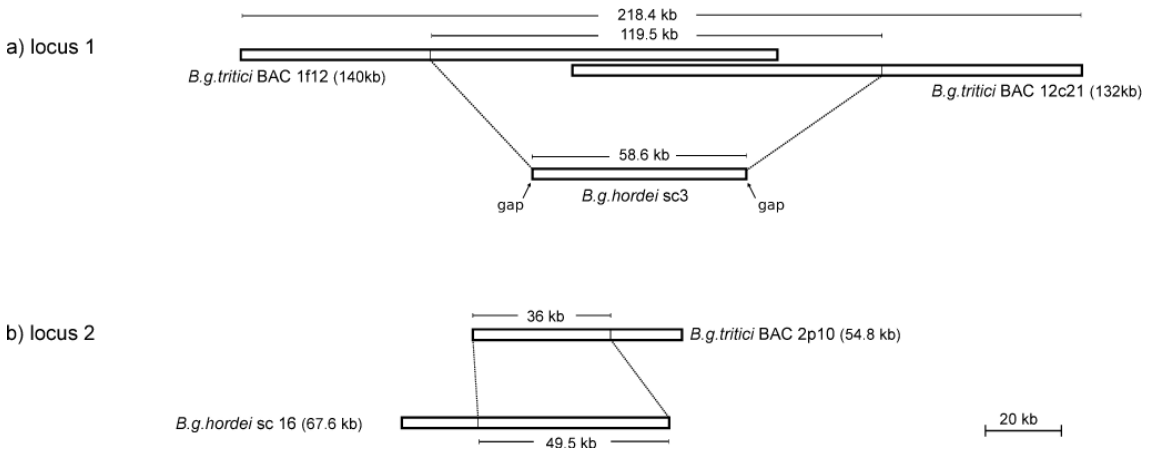

Figure 3.1. Schematic representation of locus 1 and 2 in *B.g. tritici* and *B.g. hordei*: a. locus 1 is located on the two overlapping*B.g. tritici* BACs 1f12 and 12c21 (total length 218.4 kb) and the *B.g. hordei* scaffold (sc) 3. For comparative analysis, 119.5 kb in *B.g. tritici* and 58.6 kb in *B.g. hordei* were aligned. b. Locus 2 is located on *B.g. tritici* BAC 2p10 (54.8 kb) and on *B.g. hordei* scaffold 16. An alignment of 36 kb *B.g. tritici* sequence and 49.5 kb *B.g. hordei* sequence was used for comparative analysis.

sequence revealed eight genes, while 49.8% of the total sequence was classified as transposable elements. A region of 119.5 kb in *B.g. tritici* can be aligned to 58.6 kb continuous sequence (no sequence gaps) on *B.g. hordei* scaffold 3 (Figure 3.1a). The *B.g. hordei* locus consists of 32.6% TEs and contains six genes. A detailed alignment of the annotated 119.5 kb *B.g. tritici* and the 58.6 kb *B.g. hordei* sequence (Figure 3.2) revealed microcolinearity of the terminal regions ($\sim$30 and $\sim$16 kb in *B.g. tritici*; $\sim$25 and $\sim$11 kb in *B.g. hordei*), separated by a block of non-colinear sequences of 54.9 kb in *B.g. tritici* and 11.6 kb in *B.g. hordei* (Figure 3.2). Six of the eight *B.g. tritici* genes have a homolog in the *B.g. hordei* locus, and these genes are found in colinear positions with the same transcriptional orientation. The two remaining *B.g. tritici* genes are not found on the *B.g. hordei* sequence because they are located outside of the 119.5 kb region that was aligned (Figure 3.1a). Descriptions of the six colinear genes are given in Table 3.2. The coding regions of orthologous genes are 81-95.8% identical (Supplementary Table B.1), and up-and downstream regions up to 1 kb distance to start or stop codon are slightly less conserved than the exons (Figure 3.2).

On the *B.g. tritici* locus 1, we identified 42 TEs, while on the *B.g. hordei* locus 1, we found only 21. This difference in TE content explains the large difference in size between the *B.g. hordei* and the *B.g. tritici* locus (Figure 3.2). The *B.g. tritici* sequence contains four fulllength LTR retrotransposons and six solo-LTRs (the results of unequal recombination between the LTRs of a full-length element), whereas no such elements were found in the *B.g. hordei* sequence. LTR elements altogether contribute 22.2% to the annotated sequence in *B.g. tritici*.

Two *B.g. tritici* SINE families, Yhi and Lie, consist of short elements with a length of a few hundred base pairs. The homologous families in *B.g. hordei* are named EGH24 (Rasmussen *et al.*, 1993) and EG-R1 (Wei *et al.*, 1996), respectively. Elements of these SINE families are by far the

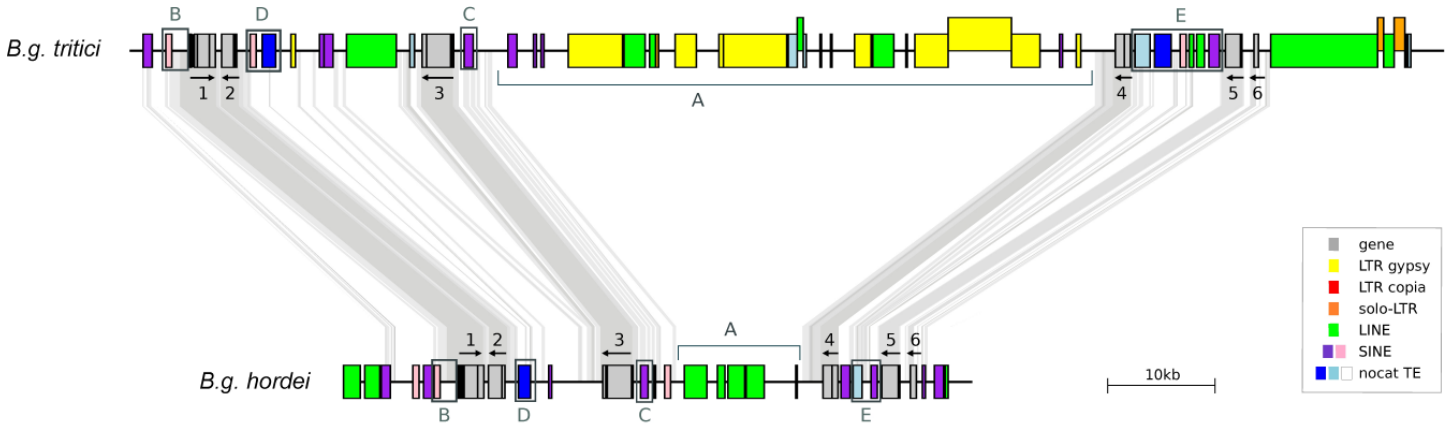

Figure 3.2. Comparison of locus 1 in *B.g. tritici* and *B.g. hordei*. Orthologous genes (gray boxes) are numbered and transcript orientation is indicated by arrows (for gene annotation see Table 3.2). Syntenic regions are shaded according to the level of conservation, with darker regions indicating higher sequence conservation. Shading levels indicate conservation of at least 95%, 90%, 85% or 80%. (A) Non-syntenic region of 54.9 kb length in *B.g. tritici* and 11.6 kb in *B.g. hordei*. (B and C) TEs conserved in *B.g. tritici* and *B.g. hordei* and located close to genes. (D) TE insertions specific for *B.g. tritici* between a gene and a conserved TE. (E) TEs insertions specific for *B.g. tritici*, between two conserved TEs. Nocat TE: unclassified TE.

Table 3.1. Contribution of TEs in the sequences aligned between *B.g. tritici* and *B.g. hordei*.

	Locus 1				Locus 2			
	B.g. tritici		*B.g. hordei*		*B.g. tritici*		*B.g. hordei*	
Superfamily	Elements	Sequence [%]	Elements	Sequence [%]	Elements	Sequence [%]	Elements	Sequence [%]
SINE[a]	12	5.7	12	11.4	11	11.4	14	12.7
LINE[b]	7	17.7	15.4	7	29.6	10	10	13.4
LTR[c] retro-transposon	10	22.2	0	0.0	0	0.0	1	0.7
soloLTR	6	1.8	0	0.0	4	2.5	6	2.5
Unclassified TE	7	4.9	2	2.0	4	4.4	6	8.1

[a] SINE: short interspersed nuclear element
[b] LINE: long interspersed nuclear element
[c] LTR: lont terminal repeat

most abundant TEs with a total copy number of 12 in each sequence (Table 3.1). However, because of their small size, their total contribution to the analyzed *B.g. tritici* and *B.g. hordei* sequences is only 5.7% and 11.4%, respectively. The two TE families Pele and Laka, which have a high copy number in both sequences (seven copies in *B.g. tritici*, two copies in *B.g. hordei*, Table 3.1) could not be classified as any known TE superfamily. Elements of the Pele family have a length of about 1450 bp and encode a transcript of unknown function, whereas elements of the Laka family have a size of only 838 bp and no ORF.

In contrast to the genes which are perfectly colinear, the intergenic regions of the *B.g. tritici* and the *B.g. hordei* sequences are much less conserved. This is due to many TEs, which have inserted at different sites in the *B.g. tritici* and *B.g. hordei* sequences after the two species diverged (Figure 3.2). Only a few TEs are found in orthologous positions (i.e. they have inserted at that position already in the common ancestor of *B.g. tritici* and *B.g. hordei*). All orthologous TEs are located close to genes, and with increasing distance from genes, fewer orthologous TEs are found (Figure 3.2). Regions B and C (Figure 3.2) show two TEs located within less than 2 kb distance to a gene and are conserved in both sequences (Figure 3.2). These elements were already present in the common ancestor of *B.g. tritici* and *B.g. hordei* and therefore represent orthologs. Conversely, additional elements have inserted either in *B.g. tritici* or *B.g. hordei* between a gene and conserved repeats. This is the case in region D where an insertion occured 900 bp away from gene 2 only in *B.g. tritici*, or in region E which has an insertion 273 bp away from gene 4 in *B.g. hordei* (Figure 3.2). We could determine the precise borders of these two elements (i.e. their precise insertion site which is flanked by a diagnostic target site duplication), demonstrating that these two insertions must have occurred after the divergence of the two species.

Drastic rearrangements seem to have taken place in the middle region of the locus since species divergence: region A separates genes 1-3 and genes 4-6 by over 55 kb in *B.g. tritici* while the cor responding region in *B.g. hordei* is only 11.6 kb (Figure 3.2). Besides solo-LTRs, SINEs and truncated LINEs, there are several nested gypsy LTR retrotransposons in *B.g. tritici*. Three of

Table 3.2. Genes in the *Blumeria* sequences and their homologs in *S. cerevisiae* and *M. oryzae*.

Gene number	Name	*B.g. tritici* protein (aa)	*B.g. hordei* protein (aa)	*S. cerevisiae* homolog	*M. oryzae* homolog	Annotation
Locus1						
1	Cac	700	700	YML102W (469 aa)	MGG_03737 (714 aa)	Component of chromatin assembly complex
2	Pat	465	467	YPL076W (280 aa)	MGG_03716 (481 aa)	Phosphatidylinositol N-acetylglucosaminyl-transferase activity
3	Rad16	884	880	YOR191W (1619 aa)	MGG_00453 (994 aa)	DNA repair protein RAD16
4	Zfp	444	444	No homolog	MGG_04317 (535 aa)	C3HC zinc finger domain-containing protein
5	Atf	506	506	YJL0910C (491 aa)	MGG_01189 (499 aa)	Acyltransferase activity (GPI anchor bio-synthesis process)
6	Hp1	164	168	YAL046C (119 aa)	MGG_09075 (162 aa)	Hypothetical protein
Locus2						
1	Ppat	387	406	YGR277C (305 aa)	MGG_08064 (415 aa)	Pantetheine-phosphate adenylyl-transferase
2	Est	249	249	YOR126C (239aa)	MGG_05726 (262 aa)	Esterase (SGNH-hydrolase type)
3	Min	336	332	YJL208C (329 aa)	MGG_05324 (334 aa)	Mitochondrial nuclease
4	Hp2	643	643	YPL210C (640 aa)	MGG_06904 (672 aa)	Hypothetical protein
5	Ccp	788	788	YJR040W (779 aa)	MGG_11357 (786 aa)	Chloride channel protein
6	Een	303	303	YDR280W (305 aa)	MGG_0860 (294 aa)	Exosome complex exonuclease

them are complete elements and therefore could be used to calculate the insertion time by determining the nucleotide differences between the two LTRs of each element (SanMiguel *et al.*, 1998; Martin *et al.*, 2010a). For two elements, insertion times were estimated at 2.3 $\pm$ 0.6 and 2.6 $\pm$ 0.6 million years ago, respectively. The third element has inserted very recently and the exact insertion time could not be determined because the two LTRs were 100% identical. Two additional LTR elements are either truncated or split by the insertion of another element. In the corresponding region of the *B.g. hordei* locus, which has a length of 11.6 kb, only three truncated LINEs and one SINE element were found (Figure 3.2).

Locus 2 has a similar size despite dynamic reshuffling of intergenic sequences

BAC 2p10 has a size of 54.8 kb and comprises the major part of the orthologous region of the second *B.g. hordei* scaffold sc16 that was studied (Figure 3.1b). The *B.g. tritici* BAC sequence contains six genes (Figure 3.3). Details on their annotation and their homologs from yeast and M. oryzae are described in Table 3.2. All annotated TEs add up to 48.2% of the entire sequence, the majority of the TEs being short elements with a size of less than 1 kb, primarily SINEs or truncated LINEs (Table 3.1). Elements of the Laka family are also relatively abundant in locus 2, with four elements within less than 30 kb (Figure 3.3, light blue boxes). Interestingly, all these elements are located in immediate proximity to genes or within less than 2 kb distance. There are no full-length LTR retrotransposons, and the four solo-LTRs are either associated with SINEs or have inserted into LINEs. Towards the 30 end of the *B.g. tritici* sequence there is an accumulation of LINE elements, which contribute 22.1% to the BAC sequence.

The orthologous *B.g. hordei* scaffold has a size of 67.6 kb. We annotated seven genes on this sequence. More than 41% of the entire sequence shows homology to TEs, the most abundant families belonging to SINEs and LINEs. In general, the TE family types and their frequency in both the *B.g. tritici* and *B.g. hordei* sequence are comparable. However, there are differences in the percentage that individual superfamilies contribute to the sequence. For example, SINE elements are twice as abundant in the *B.g. hordei* sequence compared to *B.g. tritici*, while LINE elements represent twice as much sequence in *B.g. tritici* than in *B.g. hordei*.
We could only align 36 kb of the *B.g. tritici* BAC with 49.5 kb of the *B.g. hordei* scaffold for a comparative analysis (Figure 3.3) because the orthologous regions were located at opposing ends of *B.g. tritici* BAC sequence and *B.g. hordei* gap-free region (Figure 3.1b). Consistent with our observations at locus 1, we found that the six genes which are present in the *B.g. tritici* sequence are conserved in colinear positions also in *B.g. hordei*. The corresponding coding sequences are at least 86.4% identical to each other, with a maximum identity of 94.7% for gene 6 (Supplementary Table B.1). Sequence conservation drops with increasing distance from coding sequences (Figure 3.3, details A and B), ending about 1 kb upstream of all genes except for gene 5, where about 3 kb including some TEs are mostly conserved. Similarly, highly conserved TEs can be found in the 4 kb intergenic region between genes 1 and 2 (Figure 3.3, detail C), and also upstream of gene 4 (Figure 3.3, detail F) although these orthologous elements are less than 80% identical (cutoff for shading in Figure 3.3 is 80% conservation). While genes are again all in colinear positions, there are two intergenic regions populated with TEs which are completely different in the two species. An 8.7 kb cluster in *B.g. hordei* consisting of two truncated LINEs, three SINEs and one solo-LTR is absent in *B.g. tritici* Figure 3.3, detail E), and the region between genes 2 and 3 differs completely in the *B.g. tritici* and *B.g. hordei* sequences regarding TE composition Figure 3.3, detail D). While a SINE and a unclassified element are present in this region in *B.g. tritici*, the corresponding sequence in *B.g. hordei* comprises a LINE element interrupted by multiple solo-LTRs.

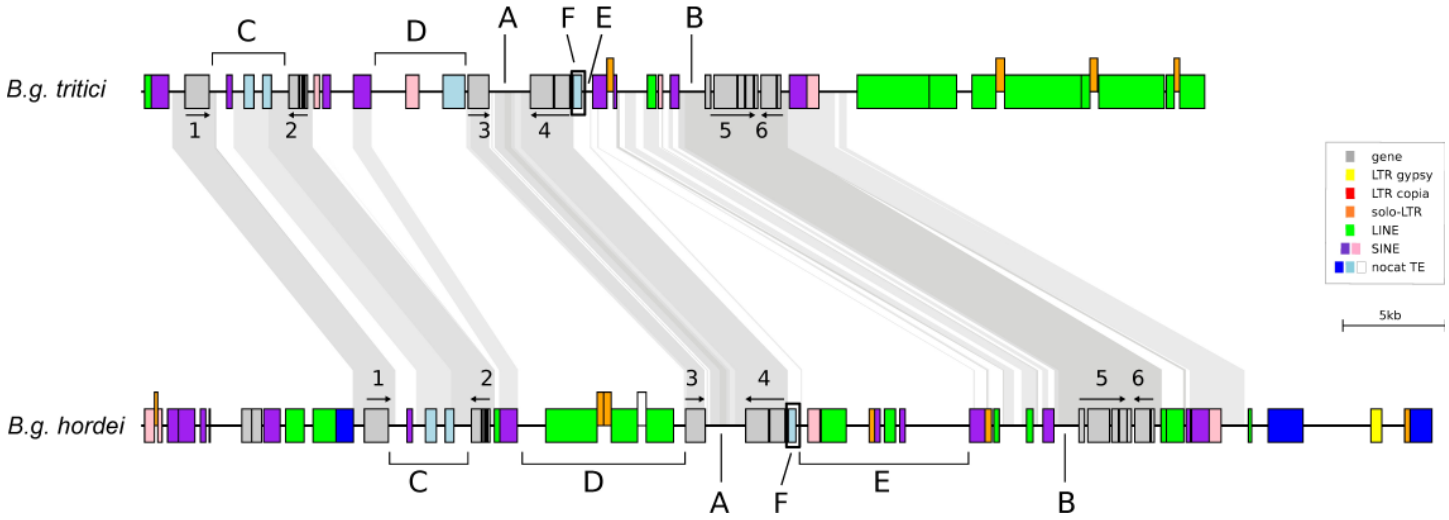

Figure 3.3. Comparison of locus 2 in *B.g. tritici* and *B.g. hordei*. Orthologous genes (gray boxes) are numbered and transcript orientation is indicated by arrows (for details see Table 3.2). Syntenic regions are shaded according to the level of conservation, with darker regions indicating higher sequence conservation. Shading levels indicate conservation of at least 95%, 90%, 85% or 80%. (A and B) Conserved non-conding sequences up- and downstream of genes. (C) Intergenic region with conserved TEs. (D and E) Non-conserved intergenic regions. (F) Conserved TE upstream of a gene. Nocat TE: unclassified TE.

B.g. tritici and *B.g. hordei* diverged approximately 10 MYA

The sequence alignment of *B.g. tritici* BAC clones with their orthologous regions in *B.g. hordei* revealed that orthologous genes, which are well conserved and colinear, are embedded in regions of high diversity caused by transposon activity. This extensive genomic divergence raised the question of the phylogenetic relationship between *B.g. tritici* and *B.g. hordei*. In the context of host-pathogen evolution, the divergence time of *B.g. tritici* and *B.g. hordei* is of particular interest. To estimate the divergence time of the two pathogens, we used intergenic sequences (preferrably TEs) as follows: if a TE sequence is conserved in orthologous loci (i.e. the TE has inserted already in the common ancestor), this sequence can be used for molecular dating because TEs are largely free from selection pressure and therefore accumulate mutations at a basic background rate (Petrov, 2001). Thus, the number of nucleotide substitutions accumulated in orthologous TEs in the two species is directly proportional to the time of divergence of the two species. If one has knowledge of the average of the background substitution rate, the number of substitutions in the orthologous TEs can be used to calculate the divergence time. The substitution rate in plant genomes has been estimated to be 1.3 x 10^{-8} nucleotide substitutions per site per year (Ma and Bennetzen, 2004). (Kasuga *et al.*, 2002) found that mutation rates in fungi, plants and animals are all very similar and, consequently, we used the mutation rate of 1.3 x 10^{-8} nucleotide substitutions per site and year. This rate has also been used for molecular dating of LTR retrotransposon insertions in the truffle genome (Martin *et al.*, 2010a).

In our analysis, we used only intergenic sequences that were at least 1 kb away from start or stop codons of genes in order to exclude non-coding sequences that might be under selection pressure (e.g. promoters and downstream elements). Using this restriction, a total of 13 kb conserved intergenic sequence containing TEs and non-coding sequence was defined, 8.5 kb from locus 1 and 4.5 kb from locus 2. The application of the basic substitution rate to the 8.5 kb and 4.5 kb of orthologous intergenic sequences resulted in divergence time estimates of 10.09 Myr ($\pm$0.25) for locus 1 and 10.16 Myr ($\pm$0.34) for locus 2 (Table 3.3).

Table 3.3. Divergence time estimates for *B.g. tritici* and *B.g. hordei*.

Locus	Sequence	Sites (bp)	Transitions	Transversions	Divergence time MYA
Locus 1	Intergenic	8556	1002	876	10.09 *pm* 0.25
Locus 2	Intergenic	4556	506	502	10.16 *pm* 0.34

3.4 Discussion

Analysis of large, contigous sequences from *B.g. tritici*

Our annotation of 273.3 kb BAC sequence from *B.g. tritici* gives an insight into the genome structure of a cereal powdery mildew. Remarkably, about 50% of the sequence is contributed by TEs. This is probably an underestimate both for this specific region as well as for the whole genome since our TE database is possibly not complete and we have specifically selected gene-containing regions. The very high percentage of TEs is comparable to what was observed in the truffle genome (∼58% repetitive DNA) and in the genome of the plant pathogenic oomycete P. infestans (∼74% repetitive DNA). TEs in the *B.g. tritici* and *B.g. hordei* sequences are mostly truncated or degenerated, and full-length copies of larger elements like LINEs or LTR retrotransposons (3-7 kb length) are rare. This indicates past and present transposon activity in the genome, and this is supported by the relatively young LTR retrotransposons we found (insertion times 1-2.6 MYA).
For annotation of genes in *B.g. tritici* and hordei sequences, we have chosen protein reference sets from yeast because of the high quality annotation and M. oryzae as a phylogenetically closer dataset. The predicted proteins in Blumeria corresponded well to M. oryzae reference proteins in terms of length and homology. Compared to yeast proteins however, there are significant differences in size in about 50% of the cases, and almost all of them occur in locus 1 (Table 3.2). We found no indication for gene clustering in the 273.3 kb that we analyzed.

Comparative analysis of *B.g. tritici* and *B.g. hordei* orthologous regions

Comparison of the *B.g. tritici* sequences with their orthologous regions in *B.g. hordei* revealed that the genes of these two loci were colinear and well conserved. The coding sequence conservation ranged from 81% to 96% nucleotide identity (peak at 93-94%; Supplementary Table B.1. Assuming that *B.g. tritici* and *B.g. hordei* diverged 10 million years ago, basic mutations would in average lead to ∼13% nucleotide difference in sequences which are not subjected to selection pressure. In contrast, many genes are under purifying selection and, therefore, accumulate mutations less frequently than non-coding sequences. This is likely to be the case for the 10 genes that are at least 87.5% identical between *B.g. tritici* and *B.g. hordei* (Supplementary Table B.1). The two remaining genes (locus 2/gene 3 (86.4%) and locus 1/gene 6 (81%) identity) are probably under weak selection pressure or under diversifying selection.
There is a high abundance of TEs in intergenic regions and their differential insertion patterns generate high sequence diversity. The genomes of *B.g. tritici* and *B.g. hordei* share a large number of homologous TE families, such that the same TE database could be used for annotation of *B.g. tritici* and *B.g. hordei* sequences. Non-orthologous members of a homologous TE family in *B.g. tritici* and a *B.g. hordei* were up to 90% identical (e.g. SINE families). Orthologous TEs, i.e. copies that inserted before the divergence of *B.g. tritici* and *B.g. hordei*, were rarely found. Instead, TEs are mostly not conserved between *B.g. tritici* and *B.g. hordei* sequences, indicating that there has been high TE activity and intense reshuffling of intergenic squences in both genomes since the divergence of *B.g. tritici* and *B.g. hordei*.

Opportunities and limitations of molecular dating in mildew fungi

The few available studies on the phylogeny of cereal powdery mildews suggest *B.g. tritici*-*B.g. hordei* divergence times of 14,000 years (Wyand and Brown, 2003), 4.6 MYA (Inuma *et al.*, 2007) and 10 million years (Takamatsu and Matsuda, 2004). The wide range of estimates does not provide conclusive information regarding the hypothesis of host-pathogen co-evolution. The basis of previous estimates were sequences like rDNA fragments/ITS sequences or the b-tubulin (tub2) and chitin synthetase genes. The use of these sequences for molecular dating is contentious because they are relatively short and quite conserved among the B. graminis ff. spp. and therefore only provide limited information. For example, the ITS sequences of *B.g. tritici* and *B.g. hordei* are 91.9% identical (Inuma *et al.*, 2007), and the divergence time of 10.6 MYA calculated on 28S rDNA (Takamatsu and Matsuda, 2004) resulted in a standard deviation of $\pm$3.6 million years because only 8 polymorphisms were observed in the 600 bp sequences that were used for comparison. In addition, it was found that the frequency of mutations in rDNA ITS regions was much lower than in the tub2 gene sequences (Wyand and Brown, 2003), although ITS regions are considered to be non-functional. Therefore, Wyand and Brown (2003) suggested to use faster evolving functional genes or rDNA IGS (internal gene spacer) regions instead of rDNA or ITS. We have chosen non-gene-coding conserved regions of two unlinked loci within the *B.g. tritici* and *B.g. hordei* genome for molecular dating. This allowed us to identify long phylogenetically informative sequences which are not subjected to any kind of selection pressure.

The use of adequate mutation rates for specific sequences is crucial to produce precise and reliable divergence time estimates. For Erysiphales, mutation rates for ITS and 28S rDNA sequences have been determined and applied (Takamatsu and Matsuda, 2004). To our knowledge, our study is the first one where non-coding, conserved sequences were used to determine phylogenetic distances in Blumeria species. No mutation rate specific for intergenic regions in Erysiphales genomes is available. However, (Kasuga *et al.*, 2002) suggested that the range of neutral mutation rates is constant among the clades of plants and fungi, and this was shown on data from six protein-coding genes and the ITS of rDNA. Therefore, we used the mutation rate for intergenic regions in plants (Ma and Bennetzen, 2004). This rate has also been used to determine major cycles of LTR retrotransposon activity in the truffle genome (Martin *et al.*, 2010a). Our estimate of 10.09 Myr ($\pm$0.25) and 10.16 Myr ($\pm$0.27) divergence time is fairly consistent with the 10.6 $\pm$ 3.6 MYA suggested by Takamatsu and Matsuda (2004), but with a much lower standard deviation. We conclude that the estimate presented in this study based on a total of 13.3 kb intergenic sequence is more robust. Once the complete genome sequence of *B.g. tritici* will be available, it can be used for even more accurate molecular dating.

The *B.g. tritici*-*B.g. hordei* divergence time of $\sim$10 MYA suggests host-pathogen co-evolution over most of the time since the divergence of wheat and barley (divergence 12 MYA). This is conflicting with Inumas' estimate based on ITS sequences of 4.6 $\pm$ 2 MYA (Inuma *et al.*, 2007) and with Wyand and Brown (2003), who suggested host-jumping and recent divergence of less than 14,000 years for several cereal mildew pathogens. The conclusions of Wyand and Brown (2003) were based on a discrepancy between the tree topology of the oat mildew pathogen and its host when put in context with the other cereal hosts and powdery mildew pathogens (Supplementary Figure B.1). Our study focused on the divergence of *B.g. tritici* and *B.g. hordei*

only and, therefore, does not provide information on the evolutionary relationship of other *Blumeria* f. sp. like the rye and oat mildew pathogens.
It is interesting to note that our estimated divergence of *B.g. tritici* and *B.g. hordei* seems to have occurred two million years after the estimated divergence of wheat and barley (Chalupska *et al.*, 2008). If both estimates are correct, the time lag of two million years could represent a phase where the pathogen was still able to infect both its closely related hosts. The period of host-range expansion would have come to an end once the hosts had evolved unique physiological properties and the pathogen had adapted to one host. After this specialization, the pathogens would then have co-evolved with their hosts. Such a combination of host-range expansions and host-shifts was suggested in a previous study (Inuma *et al.*, 2007) for the closely related *B.g. tritici*, Blumeria graminis ff. spp. tritici, agropyri and secalis (the mildew pathogens of wheat, Agropyron and rye). These three powdery mildew *formae speciales* can still mate and produce normal ascomata and ascospores. Hybridization of *B.g. hordei* and tritici is also still possible, but only ascomata are produced and formation of ascospores is rare (Inuma *et al.*, 2007). Distantly related Blumeria graminis ff. spp. like avenae and poae cannot be hybridized with triticum clade B. graminis. Thus, it seems possible that host specialization is a process that can last several million years.
Additionally, there are sometimes extreme incongruities in the phylogenetic trees of hosts and pathogens where actual host-jumps to more distantly related species must have occurred, such as the jump of a close relative of *B.g. tritici* and *B.g. hordei* from Triticeae to the more distantly related Avena species (Wyand and Brown, 2003). We conclude that the evolution of host-pathogen relationships of cereals and B. graminis ff. spp. can take many different routes. These can range from dramatic host-jumps, resulting in closely related parasites infecting distantly related hosts, over flexible host shifting to a conservative long-term co-evolution of hosts and pathogens. Future studies that include a wider range of host-parasite pairs will be necessary to determine which paths are the ones most frequently taken. Genome-wide comparative studies on *B.g. tritici* and *B.g. hordei*, and comparison to the genomes of the pea and Arabidopsis powdery mildew (*Erysiphe pisi* and *Golovinomyces orontii*) which will be available in the near future, will provide accurate divergence time estimates for all these species and will contribute to a better understanding of mildew pathogen-host evolution.

Acknowledgments

We thank Chris Ridout (John Innes Centre, Norwich) for providing SOLEXA sequence data and Bruce McDonald (ETH Zurich) for helpful discussions. This work was supported by the Swiss National Science Foundation Grant 3100A-127061/1 (BK) and an Advanced Grant of the European Research Council (Durableresistance 249996, BK).

CHAPTER 4

The wheat powdery mildew genome reveals unique evolution of an obligate biotroph

Thomas Wicker[1,a], Simone Oberhaensli[1,a], Francis Parlange[1], Jan P. Buchmann[1,8], Margarita Shatalina[1], Stefan Roffler[1], Roi Ben-David[1,9], Jaroslav Doležel[2], Hana Šimková[2], Paul Schulze-Lefert[3], Pietro D. Spanu[4], Rémy Bruggmann[5], Joëlle Amselem[6], Hadi Quesneville[6], Emiel Ver Loren van Themaat[3], Timothy Paape[7], Kentaro K. Shimizu[7] and Beat Keller[1]

[1]Institute of Plant Biology, University of Zurich, Zollikerstrasse 107, 8008 Zurich, Switzerland

[2]Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, Sokolovská 6, 77200 Olomouc, Czech Republic

[3]Department of Plant Microbe Interactions, Max-Planck Institute for Plant Breeding Research, Cologne, Germany

[4]Department of Life Sciences, Imperial College London, Imperial College Road, London SW7 2AZ, UK

[5]Department of Biology, University of Bern, Baltzerstrasse 6, 3012 Bern, Switzerland

[6]INRA, URGI, Centre de Versailles batiment 18, Route de Saint Cyr, 78026 Versailles, France

[7]Institute of Evolutionary Biology and Environmental Studies, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland

[8]Current address: MTT/BI Plant Genomics Lab, University of Helsinki, Biocentre 3, Viikinkaari 1, FIN-00790 Helsinki, Finland

[9]Current address: Department of Agronomy and Natural Resources, Institute of Plant Sciences, ARO, The Volcani Center, Bet Dagan 50250, Israel

[a]Both authors contributed equally to this work

Submitted 2013.

Supplementary figures and tables can be found in Appendix C.

Abstract

Wheat powdery mildew *Blumeria graminis forma specialis tritici* (*B.g. tritici*) is a devastating fungal pathogen for agro-ecosystems worldwide and economically the most important powdery mildew disease. The evolutionary origin of this pathogen as well as the impact of wheat domestication and the formation of hexaploid wheat as a new host 10,000 years ago are poorly understood. Here we report on comparative genome analyses of *B.g. tritici* and the barley powdery mildew *B.g. hordei*. With a size of 180 Mbp and a repeat content of 90%, the *B.g. tritici* genome is the largest and most repetitive fungal genome reported so far. We identified 602 genes which are under diversifying selection, suggesting that they act as effectors in host/pathogen interactions and are involved in host specialization. We sequenced the genomes of four *B.g. tritici* isolates from different geographical regions. The four isolates differ most notably in the presence or absence of multiple candidate effector genes. The genomes of the four isolates are mosaics of ancient and very diverse haplogroups which already existed prior to the domestication of wheat. The observed haplogroup patterns indicate that *B.g. tritici* propagates mainly asexually or through inbreeding. Furthermore, the diversity of old haplogroups in modern *B.g. tritici* isolates suggests that there was no dramatic loss of genetic diversity upon formation of the new host bread wheat. We conclude that *B.g. tritici*'s ready adaptation to a new host species was based on a highly diverse haplotype pool which provides a large genetic potential for pathogen variation.

4.1 Results and discussion

The onset of agriculture and the domestication of crops approximately 10,000 years ago resulted in drastic changes to plant pathogen environments. The genetically uniform agricultural ecosystems led either to rapid co-evolution of the pathogen with its host during domestication (host-tracking) or to the emergence of new pathogen species through host jump/shift or hybridization (Haas *et al.*, 2009; Stukenbrock *et al.*, 2011, 2012) (Text 4.2.1). For pathogens such as wheat leaf blotch *Mycosphaerella graminicola* and potato blight *Phytophthora infestans*, this process was accompanied by dramatic chromosomal changes and loss of genetic diversity (Haas *et al.*, 2009; Stukenbrock *et al.*, 2011) (Text 4.2.2).
Powdery mildews are obligate biotrophic fungi which grow and reproduce only on living hosts (Text 4.2.3). The disease occurs early in summer when haploid spores infect plants and asexually reproduce (Zhang *et al.*, 2005). Sexual reproduction of isolates of opposite mating types in late summer results in the formation of overwintering chaesmothecia (Text 4.2.4). Cereal powdery mildew *Blumeria graminis* has evolved into at least eight *formae speciales* which specifically infect one host species each (Inuma *et al.*, 2007). It is assumed that the pathogen uses an arsenal of effector proteins to infect the host (DeWit *et al.*, 2009; Spanu, 2012; Hückelhoven and Panstruga, 2011). If such an effector is recognised by the plant, it renders the pathogen avirulent and the effector gene becomes an avirulence gene (Michelmore and Meyers, 1998; Jones and Dangl, 2006). This genetic "arms-race" selects for effector changes or loss. It is postulated that the specific effector make-up determines the virulence spectrum of a particular mildew strain (DeWit *et al.*, 2009; Godfrey *et al.*, 2010; Zhang *et al.*, 2012; Pedersen *et al.*, 2012) (Text 4.2.5). Here, we wanted to study genetic diversity between and within powdery mildew *formae speciales* and explore the impact of the introduction of the new host (bread wheat) on *B.g. tritici* evolution (Text 4.2.6).

The reference genome sequence of *B.g. tritici* isolate 96224 consists of a backbone of 250 BAC contigs to which Roche/454 sequence scaffolds were anchored (Table C.1, Text 4.3.2). This allowed the analysis of genome organization at the megabase level. In total, 82 Mbp of the estimated 180 Mbp genome could be assembled because many highly repetitive sequences were collapsed or removed from the assembly (Text 4.4.1). We annotated 6,540 genes while over 90% of the genome was classified as TE sequences (Text 4.4.1). Most of the gene space is covered as 96% of the eukaryotic core genes were-full length and 98% partially present (CEGMA evaluation). Many gene families of the primary and secondary metabolism were reduced or absent as in other obligate biotrophs (Spanu *et al.*, 2010; Raffaele and Kamoun, 2012; Duplessis *et al.*, 2011; Kemen *et al.*, 2011) (Figure C.1, C.2, C.3, Text 4.4.2, 4.5.1). Fewer than 50% of the genes have homologs in yeast. In the more closely related *Botrytis cinerea*, 72% (4,731) have homologs. Almost 92% of the predicted *B.g. tritici* genes have homologs in *B.g. hordei*, indicating a very similar overall gene content of the two *formae speciales* and a large number of genes which are specific to the genus *Blumeria*. Of these *Blumeria*-specific genes, 437 encode candidate secreted effector proteins (CSEPs, Table C.2, Text 4.4.4).
Based on substitutions in synonymous sites of 5,258 closest gene homologs, we estimate that *B.g. tritici*/*B.g. hordei* diverged 6.3 ($\pm$ 1.1) million years ago (Text 4.4.3). This narrows down previous estimates which ranged from 4.7 to 10 million years (Inuma *et al.*, 2007; Oberhaensli *et al.*, 2011) and indicates that the two *formae speciales* diverged several million years ago, after

the divergence of their hosts 10-15 Myr ago (Akhunov *et al.*, 2003; Chalupska *et al.*, 2008). As a previous study (Oberhaensli *et al.*, 2011), we found gene order to be largely conserved between *B.g. tritici* and *B.g. hordei* while intergenic sequences are divergent due to TE insertions and deletions (Text 4.4.3).

Of the 5,258 *B.g. tritici*/*B.g. hordei* gene pairs, 96.6% have a ratio of non-synonymous to synonymous substitutions (dN/dS) of less than 0.5 (average 0.24). In contrast, CSEPs showed much higher dN/dS ratios at an average of 0.8, suggesting that they might be under diversifying selection (Figure 4.1). Indeed, 55 of the 77 CSEPs on which McDonald Kreitman-like tests could be performed, showed a positive direction of selection (Text 4.4.8). This *B.g. tritici* vs. *B.g. hordei* gene comparison allowed us to identify 165 novel genes which have no homologs in other fungi, lack a signal peptide but have a dN/dS ratio higher than 0.5. We propose that these genes are novel candidate effector proteins (CEPs, Text 4.4.4) that are either non-secreted, or secreted by non-conventional pathways (Nombela *et al.*, 2006). Taking CSEPs and CEPs together, *B.g. tritici* has 602 putative effector genes, 9.2% of its total gene complement (Text 4.4.4). Post-infection transcriptome data showed that expression of 99% of all CESPs and CSEPs, further supporting their potential involvement in the host/pathogen interaction.

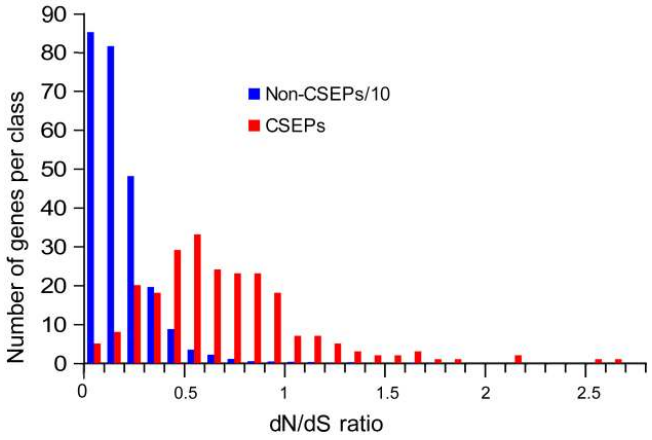

Figure 4.1. Comparison of 5,258 bi-directional closest *B.g. tritici*/*B.g. hordei* homologs. The x-axis indicates the ratio of non-synonymous to synonymous substitutions (dN/dS) for all gene pairs, the y-axis indicates the number of gene pairs in each class. The red series represents 237 gene pairs of bi-directional closest *B.g. tritici*/*B.g. hordei* homologs encoding candidate secreted effector proteins (CSEPs) while the blue series represents all other 5,021 gene pairs. For better visibility, the numbers for non-CSEPs genes were divided by 10.

In addition to the reference genome of isolate 96224 (collected 1996 in Switzerland), we sequenced isolate JIW2 (collected 1980 in England), isolate 70 (collected 1990 in Israel), and isolate 94202 (collected 1994 in Switzerland, Text 4.3.1). This allowed us to sample genetic diversity of the wheat powdery mildew gene pool in different geographical regions (from UK

and Israel) as well as within the same country (Switzerland).
The gene content of the four *B.g. tritici* isolates is almost completely identical. Besides the expected differences in the mating type locus (Figure C.8, Text 4.4.5), we identified 537 large deletions (>500 bp) in the three additional isolates. In 16 cases this led to a presence/absence polymorphisms of genes (Table 4.1). Interestingly, 13 of the 16 deleted genes are effector candidates. Considering that CEPs and CSEPs make only 9.2% of the gene content, they are highly over-represented in these presence/absence polymorphisms. CSEP analogs were described in fungal pathogens of humans and animals (Lee *et al.*, 2003; Xiao *et al.*, 2012) but specific loss of such genes has, to our knowledge, not been reported. It is possible that loss of CEP/CSEPs reflects selective pressure resulting from breeding for pathogen resistance which, unlike in animals and humans, is the norm in crop plants (Text 4.5.2).

Table 4.1. Presence/absence polymorphisms of genes in the three *B.g. tritici* isolates JIW2, 94202 and 70 compared to reference isolate 96224. Presence and absence of a gene is indicated with + and -, respectively. CSEP: candidate secreted effector protein, CEP candidate effector protein.

Gene	JIW2	94202	70	Deletion	Gene product
Bgt-3306	-	-	+	>100 kb	Mating type (Mat1-2-1)
Bgt-2805	-	-	+	>100 kb	Mating type (SLA-1)
BgtE-5545	-	+	+	44 kb	CSEP
BgtE-5597	-	-	-	25 kb	CSEP
BgtE-5802[a]	-	-	-	25 kb	CSEP
BgtE-5845	-	+	+	13 kb	CSEP
BgtE-5419	+	-	+	8 kb	CSEP
BgtE-3419	+	+	-	6 kb	CSEP
BgtAc-30466	+	-	+	5.3 bp	CSEP
BgtAc-31249	+	-	+	15 kb	CSEP
BgtAcSP-30824	+	+	-	4.5 kb	CSEP
BgtE-40100	-	+	+	1.3 kb	CSEP
BgtA-21525	-	+	-	0.6 kb	CEP
Bgt-4055	+	+	-	2.2 kb	CEP
BgtA-20784	+	+	-	9.4 kb	CEP
Bgt-369	-	+	+	13 kb	Peptidyl-prolyl isomerase
BgtAc-31336	+	-	-	0.8 kb	*ab initio*[b], no homolog
BgtA-20381	-	-	+	2.3 kb	*ab initio*[b], no homolog

[a] The two genes BgtE-5802 and BgtE-5597 are paralogs which were deleted in the same event.
[b] Gene model originates from *ab initio* gene prediction.

The 16 genes were lost in deletions spanning between 0.6 and 44 kb. Highly diagnostic sequence motifs at deletion breakpoints indicate that gene loss is the result of double-strand break repair, similar to what was described in plants (Wicker *et al.*, 2010; Buchmann *et al.*, 2012) (Figure 4.2a,

Text 4.4.6). One notable additional polymorphism was found in gene BgtE-5692 where a highly variable sequence fragment was probably introduced in a gene conversion event (Figure 4.2b, C.9). Sampling of 6 additional isolates showed no correlation between presence/absence of these genes and geographical origin of the isolates (Text 4.4.6, Table C.4).
The three re-sequenced isolates differ in 113,967 to 161,117 single nucleotide polymorphisms (SNPs) from the 96224 reference sequence, the Israeli isolate 70 being the most divergent. Small insertion and deletions of 1 to 4 bp are almost 100 times less frequent than SNPs (Table C.5, Text 4.4.7). Between 3.7-3.9% of the SNPs were found in CDS of genes and roughly 45% of them are non-synonymous. For 57% of the genes, the predicted protein was identical. In 30% of all genes, we identified two protein variants, while 10% had 3 and 3% had 4 different protein variants (Table C.6). Candidate effector genes have more non-synonymous substitutions than the average of all genes, indicating that they are under stronger diversifying selection even within the same *forma specialis* (Figure C.10, Text 4.4.7).

We observed that the SNP frequency in all isolates varies strongly in different regions of the genome compared to the reference sequence. For example, in isolate JIW2 approximately 25% of the genome consists of large segments which are nearly identical to the 96224 reference genome (0.11 SNPs/kb). These regions are distinct from regions with an approximately 10 times higher SNP frequency (Figure 4.2c and d, Text 4.4.9). This indicates that the isolates studied are mosaics of different haplogroups (i.e. chromosomal segments that are more closely related by descent than others). The average size of haplogroup segments ranges from 87.3 kb in isolate JIW2 to 150 kb in isolate 70. Based on the number of substitutions in the different haplogroup segments, we could distinguish two distinct groups representing more divergent haplogroups (H_{old}) and less divergent ones (H_{young}, Figure 4.3a). In approximately 40% of the genome we could distinguish three different H_{old} haplogroups while in about 25% of the genome, four different H_{old} haplogroups are present. In only 2.2% of the genome, all four isolates share the H_{young} haplogroup. The H_{old} haplogroups diverged approximately 43,000 to 76,000 years ago from the 96224 reference (Table C.8). In contrast, H_{young} diverged only approximately 2,100-8,600 (for isolates JIW2 and 94202) and 5,600-11,700 years ago (isolate 70) from the 96224 reference (Figure 4.3a, Table C.8).
Interestingly, the divergence of the H_{old} haplogroups coincides with the last ice age (150,000-10,000 years ago), where it is assumed that wheat ancestors were restricted to the Fertile Crescent which spans from Israel to Iran (Pinhasi *et al.*, 2005). We hypothesize that different *B.g. tritici* lineages (H1, H2 and H3 in the model in Figure 4.3b) diverged by co-evolving with different ancestral wheat populations in geographically separated areas, and that the descendants of this diversification are represented in today's H_{old} haplogroup segments (Figure 4.3b). In contrast, the H_{young} haplogroups diverged within the time period of agriculture. We speculate that northbound agricultural migration approximately 10,000 years ago could have restricted genetic exchange between European and Israeli *B.g. tritici* lineages. This would explain why the youngest haplogroup segments shared between them are fewer and diverged 8,700 (+3,000) years ago while the European isolates share haplogroups which diverged more recently (Figure 4.3b, Table C.8).

The large haplogroup segments indicate that the mildew isolates studied descended from relatively few sexual recombination events and have since reproduced mainly clonally. *B. graminis*

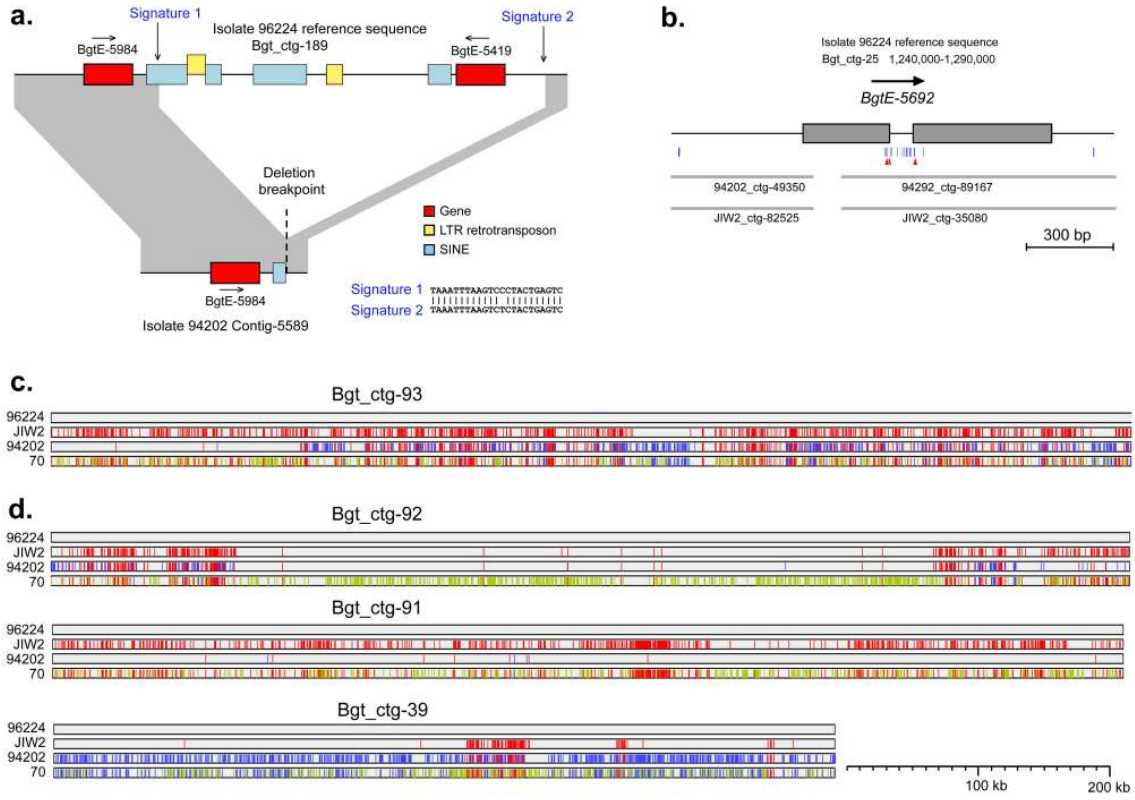

Figure 4.2. Presence/absence polymorphisms and genome sequence variation between *B.g. tritici* isolates. **a.** A map of the reference genome sequence of isolate 96224 is shown at the top. Isolate 94202 differs from the reference genome in the absence of gene BgtE-5419. The gene was lost in a deletion that removed over 8 kb. Homologous regions in the two isolates are connected with shaded areas. The presence of a nearly identical 23 bp motif (Signature 1 and 2) precisely bordering the deleted fragment indicates that the deletion is the result of a double-strand break (see Text 4.4.6, 4.5.2). **b.** The candidate effector gene BgtE-5692 contains a highly divergent segment covering parts of exons 1 and 2 (grey boxes) as well as the intron. The small size of the divergent fragment suggests that it was introduced through gene conversion. SNPs are indicated as blue vertical bars. Three SNPs which result in amino acid changes are indicated with red arrowheads. The sequence assembly of both 94202 and JIW2 contains a 110 bp gap in the 5′ region of the gene, indicating a deletion. **c.** and **d.** The *B.g. tritici* genome is a mosaic of different haplogroups. The reference genome sequence of isolate 96224 is shown at the top, underneath (in arbitrary order), the three re-sequenced isolates. Positions of SNPs are indicated with coloured vertical lines. Priority was given top-to-bottom. For example all nucleotide differences between JIW2 and the reference isolate 96224 are shown in red. If one of the other two isolates shares a SNP with JIW2, this SNP is also displayed in red. Groups of SNPs of the same colour indicate extensive chromosomal segments which originate from a different haplogroup. **c.** Large parts of the *B.g. tritici* genome are a complex mosaic of haplogroup segments that are dozens of kb long **d.** Examples for extensive regions of shared haplogroups.

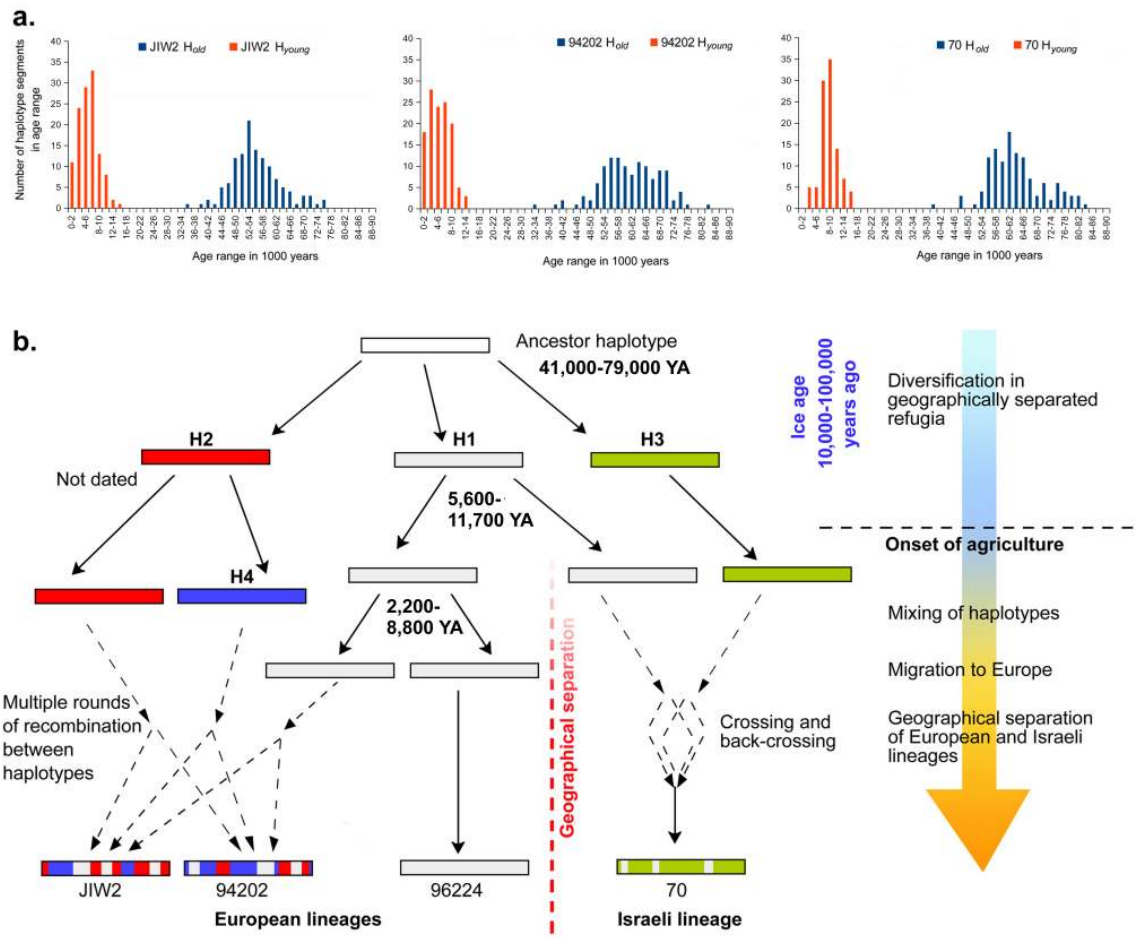

Figure 4.3. **a.** Divergence time estimates of genomic regions derived from different haplogroups. Haplogroup segments in the genomes of the three re-sequenced *B.g. tritici* isolates were divided into regions derived from a young (H_{young}) and a more ancient haplogroup (H_{old}), compared to the 96224 reference isolate. Divergence time estimate was made individually for each of the 250 FP contigs which comprise the genome. The x-axis shows ranges of divergence time estimates while the y-axis shows how many FP contigs fall in the respective age categories. Estimates for H_{young} are depicted in red and those for (H_{old}) in blue. **b.** Model for the evolution of powdery mildew isolates. The divergence and recombination of haplogroups is correlated to events such as climate change and agriculture (right side).

shows very high sexual recombination rates (Text 4.4.10). Thus, unrestricted mating of different *B.g. tritici* isolates would have completely homogenized SNP frequencies across the genomes and led to very small linkage disequilibrium (Text 4.5.3). In contrast, our observations are consistent with clonal or near-clonal reproduction (e.g. through inbreeding in small populations, Text 4.4.10, 4.5.3) which is extremely important for pathogens as it preserves successful combinations of genes and avoids acquisition of undesirable avirulence genes (Tibayrenc and Ayala, 2012; Heitman, 2006; Bougnoux *et al.*, 2008). We conclude that the distinct haplogroup patterns in the *B.g. tritici* isolates reflect strong selection for clonal propagation and/or inbreeding (Text 4.4.10, 4.5.3).

Blumeria shows a unique evolution in that it has maintained high levels of adaptability and flexibility. The genomes of *B.g. tritici* isolates are composed of haplogroup segments that predate the formation of their hexaploid bread wheat host 10,000 years ago (Salamini *et al.*, 2002). Thus, the shift from wild tetraploid to hexaploid wheat apparently has not reduced genetic diversity in *B.g. tritici* (Text 4.4.9, 4.5.4.), suggesting that the *B.g. tritici* gene pool provided all necessary genetic diversity for adaption to a range of wheat species. This is also demonstrated by its recent host range expansion to the hybrid cereal Triticale (Text 4.2.1, 4.5.4). In contrast, in *Phytophthora* (Text 4.4.11) and *Mycosphaerella*, host changes went along with the rapid formation of new species and loss of genetic diversity (Haas *et al.*, 2009; Stukenbrock *et al.*, 2011, 2012). Indeed, the youngest two *Mycosphaerella* species likely date back merely 10,000 and 500 years (Stukenbrock *et al.*, 2011, 2012) (Text 4.2.2, 4.5.4). Similarly, in *Magnaporthe oryzae*, possibly as few as three genes determine host specificity and incompatibility (Tosa *et al.*, 2006) (Text 4.5.4). This differs dramatically from powdery mildew: modern *B.g. tritici* isolates still maintain their ability to infect wild tetraploid wheat, even though the main host globally is hexaploid wheat. Additionally, *formae speciales* which diverged millions of years ago are still capable of mating (Hiura, 1978). Thus, the formation of reproductive barriers as consequence of adaptation to new hosts might be detrimental to the life style and evolutionary success of mildew.

Acknowledgements

This work was supported by an Advanced Investigator grant of the European Research Council (ERC- 2009-AdG 249996, Durableresistance), the Swiss National Science Foundation grant 310030B_144081/1 and the University Research Priority Program (URPP) Systems Biology of University of Zurich.

4.2 Supplementary text: background information

4.2.1 *Blumeria graminis* and its hosts

Powdery mildew is a disease that affects a large number of plants (Schulze-Lefert and Vogel, 2000), among them agronomically important crops like wheat and barley. The species *Blumeria graminis*, the "cereal powdery mildews", only infects grasses and includes eight so-called *formae speciales* (ff.spp.) (Inuma *et al.*, 2007). *Formae speciales* are subspecies of *B. graminis* which have specialised on one host species. The economically most important *forma specialis* (f.sp.) is powdery mildew of wheat *B. graminis* f.sp. *tritici* (hereafter referred to as *B.g. tritici*) which causes major yield losses and affects grain quality (Hsam and Zeller, 2002; Griffey, 1993). Other *formae speciales* are the powdery mildew of rye (*B.g. secalis*), oat (*B.g. avenae*), wheatgrass (*B.g. agropyri*) and barley (*B.g. hordei*). Mating between *Blumeria* ff.spp. is possible to some extent (Hiura, 1978), but only leads to progeny in specific combinations, e.g. when *B.g. tritici* and *B.g. agropyri* are crossed.
There are exceptions to this host specificity: for example strains which attack wild cereal species may have a broader host range (Eshed and Wahl, 1970). Similarly, wheat powdery mildew *B.g. tritici* has a host range that includes diploid, tetraploid and hexaploid wheat species. Occasionally host range expansions do occur, as was demonstrated when *B.g. tritici* has expanded its range within the past decade to triticale, an artificial hybrid species of wheat and rye (Troch *et al.*, 2012).

4.2.2 Evolution of plant pathogen interactions

Little is known about the evolution and genetic diversity of fungal plant pathogens. The evolution of both wheat leaf blotch *Mycosphaerella graminicola* and the potato blight *Phytophthora infestans* included host jumps followed by specialisation (Stukenbrock *et al.*, 2007, 2011; Raffaele *et al.*, 2010). *M. graminicola* is one of a group of four closely related fungal pathogens (with *M.* S1 (synonymous with *Zymoseptoria pseudotritici*), *M.* S2 and *Septoria passerinii*) which attack a wide range of grass species. *M. graminicola* arose as a new species around the time of wheat domestication about 10,000 years ago as a result of a host jump from other grasses to wheat (Stukenbrock *et al.*, 2007). Even more recently, just about 500 years ago, the new species *Z. pseudotritici* arose from the hybridisation of two *M. graminicola* haplotypes (Stukenbrock *et al.*, 2011). Similarly, *Phytophthora* jumped hosts multiple times in its recent evolution and can now infect a wide range of species including tomato and potato (Haas *et al.*, 2009). Speciation and specialisation on a new host was shown to go along with chromosomal rearrangements and rapid changes of effector and mating type genes (Stukenbrock *et al.*, 2011; Raffaele *et al.*, 2010; Haas *et al.*, 2009).
Wheat and barley powdery mildew are much more distantly related than the described *Mycosphaerella* and *Phytophthora*. They were estimated to have diverged approximately 4.7 to 10 Myr ago (Inuma *et al.*, 2007; Oberhaensli *et al.*, 2011) and have since co-evolved with their hosts. How much *B.g. tritici* and *B.g. hordei* actually differ at the genome-wide level is not known.

4.2.3 Obligate biotrophy

Powdery mildews are obligate biotrophic pathogens which can only grow on living host cells. Obligate biotrophic fungal pathogens are found in very distantly related taxa. For example, powdery mildews which are ascomycetes have diverged from rusts (e.g. *Puccinia graminis*) which are basidiomycetes probably more than 400 Myr ago (Taylor and Berbee, 2006). Because most basidiomycetes and ascomycetes are not obligate biotrophs, obligate biotrophy must have evolved multiple times independently during evolution.
Research on obligate biotrophs is hampered by the absence of cultivation on artificial media or transformation. Therefore, much of our understanding of obligate biotrophy comes from comparative analysis of genomes of obligate biotrophic with hemi-biotrophic or autotrophic fungi. Indeed, the recently sequenced genomes of barley powdery mildew (*B.g. hordei*) and the rust fungi *Puccinia graminis* and *Puccinia triticina* showed that the obligate biotrophs lack many genes coding for enzymes of the primary and secondary metabolism, carbohydrate metabolism and transporters (Spanu *et al.*, 2010; Duplessis *et al.*, 2011). Additionally, they also lost genes encoding enzymes of the metabolism of anorganic sulfate and nitrate (Spanu *et al.*, 2010; Duplessis *et al.*, 2011). Common to all obligate biotrophs are also large sets of predicted genes which encode small secreted proteins. Many of them are expressed in haustoria (Spanu *et al.*, 2010) and/or highly upregulated *in planta* (Duplessis *et al.*, 2011).

4.2.4 Sexual and asexual life cycles of *B. graminis*

B. graminis has a sexual and an asexual life cycle. The actual disease is the asexual cycle, which occurs early in summer. It begins with a haploid conidiospore landing on the leaf and penetrating the epidermal plant cell wall after formation of an appressorium (Zhang *et al.*, 2005). Inside the plant cell, the fungus establishes a highly specialised organ called haustorium which invaginates the plasma membrane of the plant cells. This close association of haustorial surface and plant plasma membrane enables the assimilation of nutrients from the plant and probably promotes the transfer of fungal components into the plant cell. After the haustorium is established, the fungus grows secondary hyphae and produces new conidiospores which are then further distributed by wind. If conditions are favourable, a colony formed from a single conidiospore can produce up to 200,000 new haploid conidiospores (Zhang *et al.*, 2005). It is the vast amount of white powdery spores produced during the asexual cycle which gave the disease its name.
The sexual cycle is initiated by dryer weather at the end of summer. Hyphae of opposite mating types (the fungal analogue of genders) fuse for a short diploid cycle. Fruiting bodies called chasmothecia are produced in which sexually produced haploid ascospores ripen. Ascospores are then the founders of a new generation in the next year. The mating type of powdery mildew is determined by a single genetic locus, the *MAT1* locus. The mating type *MAT1-1* locus contains three genes (*Mat1-1-1, Mat1-1-3* and *SLA*) while the mating type *MAT1-2* contains two (*Mat1-2-1* and *SLA*) genes (Coppin *et al.*, 1997). The sex-determining genes *Mat1-1-1, Mat1-1-3* and *MAT1-2-1* probably all evolved from a high-mobility group (HMG) transcription factor (Martin *et al.*, 2010b). Both mating types have been described for the powdery mildew of grape *Erysiphe necator* (Brewer *et al.*, 2011). In *Blumeria*, so far only the *MAT1-2* idiomorph of *B.g. hordei* has been

described (Spanu *et al.*, 2010).
Chasmothecia can remain dormant during rough weather conditions, allowing the fungus to over-winter or survive long periods of drought. Our own unpublished experiments showed that chasmothecia can be stored for years under dry condition at room temperature or at 4°C, although capacity to eject ascospores decreases with time. However, asexual conidiospores are also known to survive winter in so-called "green bridges" (Liu *et al.*, 2012). These might be isolated plants that were not harvested or small plants germinating after harvest. Early planted winter wheat can also serve as a green bridge.

4.2.5 The gene-for-gene concept of plant resistance

The complex molecular interactions between the plant and the mildew pathogen in a successful infection are poorly understood. Previous studies suggested that fungi use an arsenal of effector molecules to manipulate the plant cell (DeWit *et al.*, 2009; Spanu *et al.*, 2010). One group of putative effector genes encode candidate secreted effector proteins (CSEPs) which are thought to be secreted based on the predicted presence of signal peptides (Godfrey *et al.*, 2010; Spanu *et al.*, 2010; Zhang *et al.*, 2012). The current model proposes that such effectors can be recognised by plant resistance genes in a gene-for-gene manner (Jones and Dangl, 2006) which then triggers the plant defence reaction. If an effector is recognised by the plant, it renders the pathogen avirulent and the effector gene becomes an avirulence gene. This "genetic arms-race" between plant and fungus puts the pathogen under strong diversifying selection pressure to alter or even lose effectors once they are recognised by the plant. Within *formae speciales*, it is generally assumed that powdery mildew strains which have different virulence spectra differ mainly in the composition of their effector gene set. However, no studies on genome-wide diversity within *formae speciales* have been reported so far.
A group of powdery mildew individuals which have asexually descended from a single spore (i.e. clones) are called an isolate. Clonal propagation is known to be an important factor in the evolution of human and animal fungal pathogens (Heitman, 2006; Bougnoux *et al.*, 2008). It is assumed that it is advantageous for a pathogen to multiply clonally, because this preserves the exact combination of genes and alleles that made it successful. In contrast, sexual recombination bears the risk of losing effective virulence factor combinations or acquiring undesirable avirulence genes. Therefore, animal pathogens generally greatly reduce their sexual cycle (Heitman, 2006; Bougnoux *et al.*, 2008).

4.2.6 Evolution of *B. graminis* host species

The host species of *B. graminis* belong to the subfamily Pooideae (family Poaceae) which evolved about 20 million years ago (MYA) (Inda *et al.*, 2008). The divergence of wheat and barley occurred about 9 MYA, and rye split from wheat about 4 MYA (Huang *et al.*, 2002; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008). Modern wheat is mainly represented by bread wheat (*Triticum aestivum*, 95% of world wheat production) and pasta wheat (*Triticum durum*, remaining 5%). Both have evolved through hybridization of their diploid ancestors (Peng *et al.*, 2011): Hybridization of wild diploid wheat (*Triticum urartu*, AA genome) and a close ancestor of the goat grass *Aegilops*

speltoides (BB genome) 300'000-500'000 years ago resulted in tetraploid wild emmer (*Triticum turgidum subsp. dicoccoides*, AABB (Dvorak and Akhunov, 2005; Haudry *et al.*, 2007)). Domestication of wild emmer about 10'000 years ago in the fertile crescent of the Near East lead to a domesticated form of emmer (*Triticum turgidum subsp. dicoccum*, AABB), and later to *Triticum durum*, the pasta wheat. The outcome of a spontaneous hybridization between domesticated emmer and *Aegilops tauschii* (DD genome) 9'000 years ago was early spelt wheat (*Triticum spelta*, AABBDD) which further evolved to modern bread wheat (*Triticum aestivum*, AABBDD).

4.3 Supplementary text: material and methods

4.3.1 Choice of fungal isolates and DNA extraction

Fungal isolates and their cultivation

All *B.g. tritici* isolates used for this study were collected on *T. aestivum*. Isolate 96224 which was collected in Switzerland in 1996 was chosen as genetic source for the *B.g. tritici* reference genome. Three additional isolates were used for re-sequencing: Isolate JIW2 was collected in England in 1980, isolate 94202 was collected in Switzerland in 1994, and isolate 70 originates from Israel and was sampled in 1990 (isolates full name is Bgt#70 and is part of the collection of Prof. A. Dinoor, The Hebrew University of Jerusalem). All isolates originate from a single spore and were propagated asexually on wheat leaves as described by Srichumpa *et al.* (2005).

Extraction of DNA for Roche/454 sequencing

For 454 sequencing, conidiospores were ground with glass beads (1.7-2.0 mm) in a Mixer Mill MM300 (Retsch GmbH), then mixed with 2 ml of pre-warmed (65°C) 2x CTAB buffer (2% CTAB, 200 mM Tris/HCl pH 8.0, 20 mM EDTA, 1.4 M NaCl, 1% PVP, 0.28 M β-Mercaptoethanol) and incubated for 1h at 65°C. The volume was adjusted to 6 ml with 2x CTAB. The homogenate was extracted with an equal volume of dichloromethane : isoamylalcohol (24:1) and centrifuged for 15 min at 740 g. This step was repeated twice. RNA was digested with RNase A (10 mg/μl). DNA was precipitated with 0.7 volume of cold isopropanol and centrifuged for 10 min at 960 g. The pellet was washed for 15 min with Solution I (76% ethanol, 200 mM sodium acetate, 100 mM Tris/HCl pH 7.4), then 2 min with Solution II (76% ethanol, 10 mM NH_4 acetate) and centrifuged for 2 min at 740 g. DNA was air-dried and resuspended in 50 μl TE (10 mM Tris, 1 mM EDTA) buffer.

Extraction of DNA for Illumina re-sequencing

Conidiospores were suspended in 2 x 400 μl of Solution A (0.35 M Sorbitol, 0.1M Tris pH 7.5, 5 mM EDTA). 2.5 ml Solution B (0.2 M Tris pH 7.5, 50 mM EDTA, 2 M NaCl, 2% CTAB) plus 0.3 g sodium metabisulphite ($Na_2S_2O_5$) + 10 μl RNAse (DNAse free, 10 mg/ml) were added. The solution was dispensed as drops in liquid nitrogen and ground to a fine powder. The sample was thawed at 65°C for 15 minutes and 2 ml of chloroform were added before mixing. The sample was centrifuged for 15 minutes at 1,500 g in a swing out rotor. The supernatant was removed with a wide-bore pipette and 1 vol isopropanol was added. The mixture was centrifuged immediately for 15 minutes at 1,500 g. The pellet was dissolved in 200 μl TE buffer and then drop-dialysed on Millipore membranes (VSWP, 0.025μm pores, 25mm, cat no:VSWP02500, Millipore AG, Zug, Switzerland) over 5 liters of TE 1:10 buffer for 24 hours at 40°C under very slight agitation. Per filter, 50 μl sample were used. Sample drops were collected by pipetting. 2.6 volumes 100% ethanol and 0.1 volumes 3M sodium acetate were added and the samples chilled at -20°C overnight. Samples were collected by centrifugation (15,800 g, 20 min.) at room

temperature, washed with 70% ethanol and re-suspended in TE. Additional centrifugation steps of 5 min at 15,800 g were performed to remove clear gel (which might originate from the dialysis filter) and the supernatant containing the DNA was collected.

4.3.2 Genome sequencing

Roche/454 and Illumina genome sequencing

Genomic DNA from isolate 96224 DNA was sequenced with Roche/454 titanium technology at the Functional Genomics Center of the University Zurich (Switzerland) to approximately 13x coverage using single fragment (2,5 million reads, 900 Mbp) and 3 kb insert paired-end libraries (5 million reads, 1,653 Mbp). Illumina sequencing was performed by GATC Biotech (Konstanz, Germany, isolates 96224 and JIW2) and TGAC (Norwich, UK, isolates 94202 and 70). From each isolate, 5 μg of DNA were sequenced with paired-end libraries of 350-450 bp insert size. Isolates 96224 and JIW2 were sequenced to approximately 24-fold coverage, isolates 94202 and 70 were sequenced to approximately 50 to 70-fold coverage (Figure C.7).

Production of the reference genome sequence of *B.g.tritici* isolate 96224

Quality trimmed 454 reads were combined with 20,000 BAC end sequences (Parlange *et al.*, 2011) and assembled using Roche's Newbler assembler (version 2.5, default parameters, minimum overlap identity: 99% , minimum overlap length: 50bp). The reference genome sequences was generated by integrating the scaffolds from the 454 assembly into a BAC library fingerprint assembly, which consists of 266 contigs (hereafter called "FP contigs") with a total size of 180 Mbp (Parlange *et al.*, 2011). BAC end sequences of BACs present in the FP contigs were used as linker sequences between 454 scaffolds and FP contigs. The scaffolds were used as queries in Blast searches against a database of all BAC end sequences. To avoid random anchoring of scaffolds to repetitive DNA in BAC end sequences, we used three different stringency levels (from very stringent to less stringent) for the Blast searches. Sequence space between anchored scaffolds was filled with strings of Ns of a length estimated based on the FP contigs. The BAC end sequences of 16 short FP contigs were all repetitive and therefore they could not be used to anchor any 454 scaffolds.
Illumina sequences from isolate 96224 were used to correct the reference sequence for 454 specific sequencing errors. About 47.9 million reads (two runs on the same 350 bp insert paired-end library, read size 96 bp, 4.3/4.6 Gbp sequence data) were quality trimmed and aligned to the reference using CLC assembly cell version 3.2 (CLC bio, Aarhus, Denmark) using the program clc_ref_assemble_long with parameters -s 0.98 -l 0.95. Nucleotide differences which were present in all the aligned Illumina reads and had a minimal coverage of 2x were accepted as sequencing errors and corrected in the reference sequence accordingly.

4.3.3 Genome annotation

Identification of transposable elements

We used a combination of two *B.g. tritici* specific repeat libraries to annotate transposable elements. One library contains nucleotide and protein sequences of manually annotated transposons from *B.g. tritici* and *B.g. hordei* (Parlange *et al.*, 2011). In order to complete this library, we used the REPET TEdenovo pipeline (Flutre *et al.*, 2011) for *de novo* detection of TEs. About 2,000 consensus nucleotide sequences were generated and classified into the main orders (Class I retrotransposons : LTR, LINE, SINE; Class II DNA transposons: TIR, MITE) or were assigned to the class "unclassified". These two libraries were used by the TEannot pipeline (Quesneville *et al.*, 2005) to annotate full-length, degenerated and nested TE copies in the genome.

Gene annotation

Gene prediction in the *B.g. tritici* reference sequence was done using two approaches. Conserved genes between *B.g. tritici* and *B.g. hordei* were identified by mapping the published *B.g. hordei* genes (Spanu *et al.*, 2010) on the *B.g. tritici* reference sequence using GMAP (Wu and Watanabe, 2005). The recently sequenced *B.g. hordei* genome contains 5,854 annotated genes (Spanu *et al.*, 2010). Prior to mapping, the *B.g. hordei* gene set was carefully searched for sequences with homology to TEs or TE-related sequences (e.g. EKA homologs (Spanu *et al.*, 2010)) by running Blast searches of all *B.g. hordei* genes against an updated version of our *Blumeria* repeat database (Parlange *et al.*, 2011) which currently contains the predicted protein sequences of 74 TE families. Based on this analysis, 124 *B.g. hordei* genes were removed from the original *B.g. hordei* gene set. The remaining 5,730 *B.g. hordei* genes were mapped to the *B.g. tritici* sequence using GMAP which resulted in 5,398 *B.g. tritici* gene models. Subsequently, the identified gene models and TEs were masked on the scaffolds of the 454 assembly. The Augustus gene prediction software (Stanke and Waack, 2003) was ran on the masked sequences after it was trained on 3,143 CDS of identified *B.g. tritici* genes. *Ab initio* gene models which had homology to TEs in our repeat library were discarded, and the remaining models were mapped to the genome draft. In a final step, the structure and location of all genes including the *ab initio* models were visualized on the genome draft using IGV (integrated genome viewer, `broadinstitute.org/igv`) for manual curation.
To assign functions to gene models, we performed gene ontology analysis (GO) with Blast2Go software (Conesa *et al.*, 2005) with the entire gene set using the default settings. In addition, we performed a Blast of the protein sequences against the PFAM database and the *Botrytis cinerea* genes (`www.broadinstitute.org`). Blast searches were performed with the BLASTALL program (ncbi.nlm.nih.gov) on local Linux servers with local databases. For all analyses, Blast hits with E-values smaller than 10E-10 were considered significant. We combined all the information available to provide detailed annotation in the definition line of each gene in the fasta file.
CEGMA (Parra *et al.*, 2007) evaluation was run on the 454 scaffolds using CEGMA version v2.4.010312. CEGMA (Core Eukaryotic Genes Mapping Approach, `korflab.ucdavis.edu/Datasets/cegma/`) uses a reference set of conserved protein families that occur in a wide range

of eukaryotes. The degree by which the gene set of a genome covers the CEGMA reference set is a measure of how complete the gene space of the genome is covered.

Transcriptome sequencing

RNA was extracted from *B.g. tritici* infected wheat leaves at 4h, 8h, 12h, 24h and 48h post infection. Equal amounts of RNA from each time point were mixed and sequenced with Illumina. About 1,109 million reads (50bp read length) which represent fungal and wheat RNA from all five time points were pooled and mapped to the genome of isolate 96224. The mapping was performed with CLC genomics workbench version 6.0.1 thereby allowing only 1 mismatch per read and counting only reads which mapped to exons. A total of 7,442,144 reads (0.6%) could be mapped to the powdery mildew genome. This was expected, because most of the extracted RNA comes from wheat. For each gene, the number of reads per gene and the average coverage (total reads in bp divided by the exon length in bp) was calculated to obtain a rough estimate of the overall expression level.

4.3.4 Comparison of obligate biotrophic and non-obligate biotrophic fungi

For comparison of gene content of obligate biotrophs with other fungi, the genomes and predicted protein sequences of *Puccinia graminis* (biotroph), *Aspergillus nidulans*, *Botrytis cinerea* and *Magnaporthe grisea* were downloaded from `broadinstitute.com`. Note that the number of predicted *P. graminis* genes (20,565) in the BROAD dataset differs from the one published (Duplessis *et al.*, 2011) (17,773) due to different versions of the genome annotation. Homologs were identified at the protein level using Blastp (predicted protein against predicted protein) or tBlastn (for the identification of un-annotated genes). For all analyses, Blast hits with E-values smaller than 10E-10 were considered significant. For gene classification, sequences of PFAM domains were obtained from `http://pfam.sanger.ac.uk/` and were compiled into 9,317 consensus sequences using an original Perl script. The predicted protein sequences of all genomes used in this study were then used as query sequences in Blastp searches against the PFAM domain consensus sequences.
For genome-wide comparative analysis of metabolic pathways, we used enzymes of *Saccharomyces cerevisiae* biochemical pathways as a reference dataset (SRI International Pathway Tools, Version 14.0, `bioinformatics.ai.sri.com`). This resource integrates a total of 549 different enzymes, which were used in tBlastn searches against the genomes of *B.g. tritici*, *B.g. hordei*, *B. cinerea*, *M. grisea*, *P. graminis* and *P. triticina*.

4.3.5 Comparative analysis of *B.g. tritici* and *B.g. hordei* gene homologs

For divergence time estimates and dN/dS analysis, we determined a set of bi-directional closest homologs of *B.g. tritici* and *B.g. hordei* genes: all predicted *B.g. tritici* proteins were used in Blastp searches against the predicted *B.g. hordei* proteins and *vice versa*. Gene pairs which had each other as top blast hit in the other dataset were considered bi-directional closest homologs.

The two predicted proteins of bi-directional closest homologs were aligned with the program Water from the EMBOSS package (`http://emboss.sourceforge.net/`). To assure a good quality of the alignments, only those alignments which had a minimum of 45 aligned amino acids were used for further analysis. On average, the proteins were 92.7% identical and 96% similar. Distribution of sequence identities was narrow as 96% of all pairs are at least 80% identical. This high degree of conservation allowed robust alignments in most sequences: in 79% of the alignments, the entire *B.g. tritici* protein could be aligned with its *B.g. hordei* homolog. In 97.1%, between 95 and 100% of the proteins could be aligned. Visual inspection of a random sample of 100 alignments (50 CSEPs and 50 non-CSEPs) confirmed that protein sequences were generally very well aligned.
The protein alignment was then used as anchor to generate the corresponding CDS alignment (i.e. a new alignment was produced with the codons corresponding to each amino acid from the two sequences) to assure that indeed corresponding bases were aligned. This is an essential requirement for both molecular dating and dN/dS analyses (see below). The construction of the CDS alignment from the protein alignment was done with an original Perl script.

dN/dS analysis and tests for direction of selection

The aligned CDS of bi-directional closest homologs of *B.g. tritici* and *B.g. hordei* (see above) were processed with the yn00 program of the PAML package (`abacus.gene.ucl.ac.uk/software/paml.html`) (Yang, 2007). yn00 implements the method of Yang and Nielsen (Yang and Nielsen, 2000) for estimating synonymous and non-synonymous substitution rates. For each of the gene pairs, the dS rate (synonymous substitutions per synonymous site) and dN rate (non-synonymous substitutions per non-synonymous site) was calculated. The dN/dS ratios were assessed separately for the 5,021 non-CSEP and 237 CSEP gene pairs to test whether the group of CSEPs shows characteristics of positive selection.
dS of all non-CSEPs and CSEP genes were compared to test whether some of the bi-directional closest homologs might represent deep paralogs. Here, the distribution of dS rates in the 5,021 non-CSEP alignments was used as reference with which the dS rates of CSEPs were compared (Figure C.6). We used the dN/dS ratio as a new criterion to identify novel classes of effector gene candidates. We chose the cutoff of a dN/dS ratio of 0.5 for the following reasons: First, 96.6% of the non-CSEP genes have dN/dS values smaller than 0.5. This was chosen somewhat analogous to a p-value cutoff of 0.05, separating genes into "typical" and "atypical" ones. Second, the dN/dS distribution of CSEPs has its peak at 0.5 (Figure 4.1). Therefore, this value can be viewed as the expected dN/dS of a given effector candidate. Third, the average dN/dS of CSEPs is 0.8 while the average dN/dS of CSEPs is 0.24, the average between the two is 0.52.
McDonald-Kreitman like tests (McDonald and Kreitman, 1991) were employed to estimate the proportion of adaptive substitutions and the direction of selection in CSEPs. We selected those bi-directional *B.g. tritici/B.g. hordei* CSEP homologs for which we had complete sequences for all four *B.g. tritici* isolates.

Characterisation of gene families

Candidate effector proteins (CEPs) and candidate secreted effector proteins (CSEPs) from *B.g. tritici* and *B.g. hordei* were pooled into one Blast database against which all individual sequences were used as queries. Blast outputs of all CEPs were screened for Blast hits with E-values <E-20. All proteins that showed homologies with each other at this level were pooled into a family. Phylogenetic trees using the phylip package (`evolution.genetics.washington.edu/phylip`) were drawn for all families with more than 10 members to ascertain that the family members indeed all showed significant homology with each other. If major branchings of the phylogenetic tree showed bootstrap values below 50, the families were broken up into smaller groups at these branchings.

4.3.6 Molecular dating and divergence time estimates

Choice of the synonymous substitution rate

All divergence time estimates were done on sequences which are presumably free from selection pressure namely intergenic/repetitive sequences or synonymous sites in the CDS of genes. For our calculations, we assumed that all these sequences accumulate mutations at the same rate (i.e. the synonymous substitution rate). We used a rate of 1.3E-8 ($\pm$ 2.29E-9) substitutions per site per year which was originally proposed by Ma and Bennetzen (Ma and Bennetzen, 2004) for intergenic regions in grasses. Recent studies showed that this rate could also be applied to intergenic sequences in powdery mildew because the obtained divergence time estimates were highly consistent with those of other dating methods that were based, for example, on ribosomal DNA sequences (Oberhaensli *et al.*, 2011; Takamatsu, 2004). The error estimate ($\pm$ 2.29 E-9) of the substitution rate is derived from an earlier publication by Gaut et al. (Gaut *et al.*, 1996) on which Ma and Bennetzen's rate is based (Ma and Bennetzen, 2004). Because the 1.3E-8 substitution rate is twice the rate of 6.5E-9 proposed by Gaut, the error rate of Gaut was also multiplied by two, resulting in the described error of $\pm$ 2.29 E-9.

Divergence time estimate of *B.g. tritici* and *B.g. hordei*

To estimate the divergence time of *B.g. tritici* and *B.g. hordei* we used synonymous sites in the CDS of bi-directional closest homologs. We used only alignment positions corresponding to the third codon base of codons for Ala, Gly, Leu, Pro, Arg, Ser, Thr and Val. For Leu, Arg and Ser (which all have 6 possible codons), we used only the codons starting with CT, TC and CG, respectively. These are the codons in which the third base can be exchanged without causing an amino acid change. We concatenated all the synonymous sites into one alignment and applied the 1.3E-8 substitution rate to obtain a single estimate for the whole genome. The error estimate of $\pm$ 2.29 E-9 for the substitution rate was then applied on the calculated divergence time. The Kimura two-parameter criterion was applied to weigh the transition to transversion ratio as previously described (SanMiguel *et al.*, 1998).

Molecular dating of *B.g. tritici* haplogroups

The genomic segments assigned to haplogroups H_{old} and H_{young} were used for molecular dating. For dating, genes plus 1 kb of up- and downstream regions were removed to avoid sequences which are under selection pressure. For the calculation of divergence times we used the same synonymous substitution described above (1.3 E-8 $\pm$ 2.29 E-9 substitutions per site per year, (Ma and Bennetzen, 2004)). To obtain an estimate for variance and standard deviation, haplogroup data were processed individually for each of the 250 FP contigs. For example, FP contig Bgt_ctg-2 has a size of 898 kb of non-N bases. In isolate JIW2, it contains 6 segments that correspond to haplogroup H_{old}. These 6 segments add up to 527 kb (59% of the FP contig) and they contain a total of 729 substitutions. From these numbers, two estimates for the divergence time of H_{old} from the 96224 were derived, one with a substitution rate of 1.071E-9 and one with 1.529E-9. This was done to factor in the error of the substitution rate (see above). In this case, this would result in two estimates of 43,800 and 62,600 years, respectively. The distribution of the individual divergence estimates for all FP contigs was used to calculate the overall standard deviation of the age estimate of the respective haplogroup. Here, the variance was calculated as the square of the sum of all the differences from the average ($\sum(X_i - X_{average})^2$). The standard deviation is the square root of of the variance.

4.3.7 Re-sequencing of *B.g. tritici* isolates

In addition to 96224, three more isolates were sequenced by Illumina to 24x (JIW), 52x (70) or 70x (94202) coverage, respectively (Figure C.7). The Illumina reads of the three isolates were mapped to the 96224 reference sequence with the CLC assembly cell (CLC bio, Aarhus, Denmark) to identify single nucleotide polymorphisms (SNPs). Reads were mapped allowing a maximal number of four mismatches per 96 bp (parameters clc_ref_assemble_long -s 0.95 -l 0.95). The coverage of single-copy regions (which include practically all genes) was approximately as expected between 24x (JIW2) and 70x (94202). Sequence coverage shows a gaussian distribution with the peak of the curve very close the expected coverage ("peak coverage", Figure C.7).
For targeted analysis of specific gene groups, we were granted access to Illumina sequences of 6 additional wheat powdery mildew isolates (Table C.4). These data were produced in the framework of a different research project and will be published elsewhere. In these additional genome sequences, we specifically surveyed those genes that showed presence/absence polymorphisms in the three re-sequenced isolates described in this study (Table 4.1). Additionally, we randomly selected 25 further CSEP genes and 25 non-CSEP genes. The CDS of these 50 genes were extracted from the 6 additional isolates and aligned using ClustalW.

Identification of high-confidence SNPs

Because of the high repeat content of the *B.g. tritici* genome, we applied very stringent criteria for SNP calling and quality control (see methods, Table C.5). In general, we required that SNPs and InDels are supported by at least 90% of the Illumina reads covering them. If the total sequence coverage of the position containing the polymorphism was less than 10, we required

that all reads support the SNP. Furthermore, we discarded SNPs with a coverage more than 3-fold higher than the peak coverage of the respective isolate. This was done to discard false positives in repetitive regions where Illumina reads from multiple different copies of the repeat were mapped to a single locus. Consequently, regions which showed numerous ambiguous SNPs (i.e. several Illumina reads with different bases in a specific position) were discarded in our SNP calling. This raised the concern that regions with many discarded SNPs would be mistaken for SNP-poor or SNP-free regions. Thus we plotted the number of discarded SNPs against the number of accepted SNPs in the whole genome (examples are shown in Figure C.12). We found no instances where regions with few or no high-confidence SNPs coincided with regions with many discarded SNPs. We concluded that our SNP mapping methodology provides robust results for the subsequent gene variation and haplogroup analysis.

Detection of large deletions

After the mapping process, we produced for each isolate a table which describes the coverage of each base position in the reference genome with Illumina reads from the respective isolates (using CLC assembly cell, assembly_info -d option). We searched these tables for regions larger than 500 bp where one or more isolates had a coverage of zero. Additionally, we produced *de novo* assemblies of all three isolates with untrimmed Illumina reads using CLC assembly cell clc_novo_assemble at default parameters. Assembly statistics are described in Table C.10. These sequence contigs were used for detailed comparison of specific genomic loci where deletions were detected.

Characterisation of mating type loci and crossing experiments

The mating type loci in the 96224 reference sequence and the Illumina assembly of isolate 70 were identified by Blastn search of a specific 225 bp marker for the *B.g. tritici MAT1-2-1* gene (Brewer *et al.*, 2011). In both isolates, the identified locus aligns with 100% nucleotide identity to intron 2 and parts of the adjacent exons of *MAT1-2-1*. *MAT1-1-1* was identified by a Blast search of a 126 bp *MAT1-1-1* marker (Brewer *et al.*, 2011) against the JIW2 and 94202 Illumina *de novo* assemblies. The coding sequence of *MAT1-1-1* was annotated based on homology to the MAT1-1-1 protein sequence of *E. necator*. *MAT1-1-3* was identified and annotated based on homology to *MAT1-1-3* in other ascomycetes.

Crossing experiments with *B.g. tritici* isolates

Crosses between *B.g. tritici* isolates were done as follows: parental isolates were propagated on the susceptible wheat line Kanzler. After 10-12 days, spores were collected and mixed with a paint brush. The mixture was then inoculated by brushing on the leaves of three weeks-old susceptible plants. A transparent plastic cylinder was placed on the infected plants, surmounted by a paper bag to allow airflow but prevent from any contamination by external spores. As a control to insure that no cross contamination occurred, additional pots were added to the assay: (1) pots with Kanzler plants inoculated exclusively with each one of the parental isolates, and

(2) pots with non-infected Kanzler plants. The pots were placed in the dark for 12 hours and then in a growing chamber at light/temperature condition of: Day 16h 22 °C Night 8h 18 °C. After three to four weeks during which both isolates colonized the plants, watering was stopped because drought stress of the plant triggers the fungus to enter the sexual reproductive phase. Three weeks later, chasmothecia, which are the fruiting bodies in which the meiosis occurred and which contain the ascopores, were visible as black dots on the dried leaves. The plants were dried out for two additional weeks and leaves bearing the chasmothecia were cut and placed for one month at 4°C. To proceed with the ejection of the ascospores, chasmothecia were placed in high humidity conditions on a wet Wattman paper (16th day, 15-17°C) and above fresh Kanzler leaves which were changed every day to collect the ejected spores. Single colonies were isolated to make sure that each colony corresponds to only one individual spore.

4.3.8 Automated identification of *B.g. tritici* haplogroup segments

Positions of all SNPs in the three re-sequenced isolates were mapped on the 96224 reference genome sequence and visualised with an original Perl script (Figure 4.2). The genomes of the three re-sequenced isolates are mosaics of segments which are practically identical to the 96224 reference sequence (referred to as H_{young}) and regions which have a roughly 5 to 10-fold higher SNP density (referred to as H_{old}).
The identification of the different haplogroup segments was automated as follows: SNP distribution was surveyed in sliding windows of 20 kb across the genome. Because the genome sequence contains large gaps caused by sequence scaffolds that were anchored on opposite ends of a BAC clone, sequence gaps larger than 2,000 bp were excised from the genome sequence and replaced by stretches of 200 Ns for the analysis. This analysis was done on the 128 largest FP contigs which contain at least 200 kb of non-gap sequence (i.e. 10 times the size of the sliding window). The resulting SNP density distribution is an overlay of the densities of the SNP-rich and SNP-poor regions. For all three re-sequenced isolates, density in SNP-rich regions peaked at approximately 22 SNPs per 20 kb (i.e. 1.1 SNP per kb, example in Figure C.11). For simplicity we divided the genome in segments with average SNP densities of 22 SNPs per 20 kb or higher (i.e. H_{old}) and segments with a lower SNP density. To determine a suitable cutoff between the two, we simulated SNP densities assuming a random distribution of SNPs at an average density of 22 SNPs per 20 kb. This simulation showed that practically no segments with 9 or fewer SNPs per 20 kb (i.e. approximately 1 SNP every 2,300 bp) can be expected by chance. Thus segments with lower SNP density were defined as H_{young}.
Using this cutoff value, we used distances between neighbouring SNPs to identify breakpoints between H_{old} and H_{young} haplogroups. Regions containing SNPs that were spaced at distances of less than 2,300 bp were assigned to haplogroup H_{young}. Single incidents of too closely spaced SNPs (for H_{young}) or too widely spaced SNPs (for H_{young}) were ignored. A single large spacing in a SNP-rich H_{young} region could, for example, be caused by a gap in a sequence scaffold (454 scaffold may contain gaps of few hundred bp up to 2000 bp due to linking of paired-end reads). Likewise, two SNPs could be closely spaced by chance in an otherwise SNP-poor region.
The mapping of haplogroup segments resulted in a table with start and end positions of H_{old} and H_{young} haplogroup segments for each of the four isolates. These were then used for pairwise

comparisons to determine the genomic regions where isolates share the same haplogroup and in which segments they differ. Tables with haplogroup positions for all isolates can be obtained via FTP upon request.

4.3.9 Simulations on unrestricted mating and inbreeding

We wrote perl programs that simulate the outcomes of inbreeding vs. unrestricted mating in haploid organisms with two mating types. In particular, we wanted to calculate the numbers of haplogroup (i.e. recombination) breakpoints that can be expected if populations of different haplotypes mate. The starting point of the simulation was a population of user-defined size which consists of 50% haplotype 1 and 50% haplotype 2. Each of the two sub-populations consists to 50% each of two mating types. Thus, the minimal population size is 4. The adjustable parameters are the number of recombinations in each progeny (i.e. the length of the genetic map in Morgan), the number of generations and the population size.
All Perl programs used in this study are available upon request.

4.4 Supplementary text: results

4.4.1 The *B.g. tritici* isolate 96224 reference genome sequence

Roche/454 sequencing resulted in 11,330,743 reads with an average read length of 206 bp. This provided an approximately 13-fold coverage of the 180 Mbp genome of *B.g. tritici* (Parlange *et al.*, 2011). The resulting sequence assembly has a total size of 97.4 Mbp (82 Mbp non-N bases) and consists of 3,522 scaffolds (Table C.1). The discrepancy between assembly size (97.4) Mb and actual genome size (estimated 180 Mbp, (Parlange *et al.*, 2011)) is due to the high repeat content of the genome. A large number of contigs in the assembly represent collapsed repeat sequences which is indicated by a much higher read coverage than 13x. Furthermore, about 9.6% of the 454 shotgun reads were excluded by the Newbler assembly software because they were too repetitive to be integrated.
In total 1,907 sequence scaffolds representing 67 Mbp (82%) of the assembly were anchored to a backbone of 250 BAC contigs via BAC end sequences. About 15 Mbp of mostly small sequence scaffolds which could not be anchored were pooled in an additional pseudomolecule (Bgt_ctg-10,000). The size of large sequence gaps was estimated based on the BAC FP assembly. The total size of the genome draft based on 454 assembly and BAC end sequence anchoring is 126 Mbp, of which 82 Mbp are non-N bases (Table C.1).

B.g. tritici has at least five chromosomes

Estimates from chromosome counting or linkage maps are not available for *B.g. tritici*. To our knowledge, the sequences of centromeric repeats in powdery mildew are still unknown. It was therefore not possible to determine the number of chromosomes based on the numbers of centromers. *B.g. tritici* telomeres apparently consist of tandem repeats of TTAGGG motifs, as was also found in other filamentous fungi (Javerzat *et al.*, 1993). We identified such repeat arrays on 10 sequence scaffolds, indicating the existence of at least 5 chromosomes. However, this figure must be taken with caution, as coverage of chromosome ends is very poor in whole-genome shotgun sequencing and technically impossible in the BAC approach. Nevertheless, these data agree with observations of (Borbye *et al.*, 1992) who postulated five small and at least two large chromosomes in *B.g. hordei*.

The *B.g. tritici* genome contains more than 90% repetitive DNA

We previously estimated 85% of the *B.g. tritici* genome to consist of repetitive DNA and transposable elements (Parlange *et al.*, 2011). Genome wide annotation of TEs using REPET TEannot pipeline confirmed that over 75% of the 82 Mbp non-N bases are repetitive. The actual repeat content of the *B.g. tritici* genome however is significantly higher. On one hand we know from assembly statistics that approximately 9.6% of the 454 reads could not be assembled due to their highly repetitive nature and we expect that many repetitive sequences were collapsed into consensus contigs during the assembly. On the other hand, both the Newbler assembly software as well as total size of the FP BAC contigs indicate a genome size of approximately 180 Mbp. We

therefore estimate that the total repeat content of the *B.g. tritici* genome is at least 90%, which would be substantially higher than the 75% TEs reported for the *B.g. hordei* genome (Spanu *et al.*, 2010). In any case, the repeat content of *B.g. tritici* is much higher than in the genomes of *Tuber melanosporum* (58% TEs, 125Mbp, (Martin *et al.*, 2010a)), *Puccinia graminis* (45% (Duplessis *et al.*, 2011)), *Botrytis cinerea* (3-4% (Amselem *et al.*, 2011)) and *Magnaporthe grisea* (10% (Dean *et al.*, 2005)).

B.g. tritici has a gene number similar to *B.g. hordei* and yeast

The *B.g. tritici* gene set includes 5,398 genes which were identified through homology with *B.g. hordei* genes and 1,142 *ab initio* gene models, resulting in a gene number of 6,540. The gene space is mostly covered by the assembly because CEGMA evaluation (Core Eukaryotic Genes Mapping Approach (Parra *et al.*, 2007)) showed that 237 (95.56%) out of 248 Core Eukaryotic Orthologous Groups (KOGs) were-full length and 244 (98.39%) partially present. This indicates that, despite the obviously incomplete coverage of the repetitive fraction, the whole-genome shotgun approach succeeded in covering the gene space almost completely. The CEGMA evaluation results for *B.g. tritici* are comparable or even higher than those of other sequenced genomes of biotrophic fungi (Kemen and Jones, 2012) (*B.g. hordei*: 94.0% covered full-length, 96.0% covered partially). One of the 248 KOGs (SOD1/KOG0441) is known to be missing in powdery mildew genomes (Spanu *et al.*, 2010).
The total gene number of 6,540 is comparable to the approximately 5,700 genes that were annotated in yeast (broadinstitute.org), but low in comparison to other fungal genomes which usually contain between 10,000 and 20,000 genes (Amselem *et al.*, 2011) (Table C.9). Although the *B.g. tritici* gene number is similar to that of yeast, only 3,182 (48.6%) of the predicted *B.g. tritici* genes have yeast homologs (Table C.2). This is very similar to *B.g. hordei* where 3,254 genes have yeast homologs (Spanu *et al.*, 2010). Of the *B.g. tritici* genes which have no yeast homologs, 205 have homology to a PFAM domain, and another 1,344 have a homolog in *Botrytis cinerea* (Table C.2).
Gene annotation is difficult and gene number estimates have to be taken with caution (Bennetzen *et al.*, 2004). Therefore, the 1,142 *ab initio* gene models were closely inspected. A total of 775 models show homology to genes in other fungal genomes and 425 of them also have a *B.g. hordei* homolog, which supports these models. Therefore, 6,173 out of the total 6,540 are high-confidence gene models.

4.4.2 Comparison of gene complements from obligate and non-obligate biotrophic fungi

To study whether gene content can be associated with the fungus' biotrophic lifestyle, we compared the gene complements of mildews and rusts (*B.g. tritici* and *Puccinia graminis*, both obligate biotrophs) with those of *Aspergillus nidulans* (oligotroph), *B. cinerea* and *M. grisea* (hemibiotrophs). While the total number of predicted genes varies greatly among the five genomes, the number of genes that show homology to the PFAM protein domains is similar in all of them (2,010-3,382, Table C.9). Interestingly, the number of different PFAM domains varies even less among

the five genomes: all fungi possess genes with homology to 1,079-1,439 different PFAM domains (Table C.9). This indicates that similar sets of gene families are represented in all five genomes. However, the fact that the number of identified PFAM families is so similar in all five fungal genomes while the number of predicted genes ranges from roughly 6,000 (*B.g. tritici* and *B.g. hordei*) to over 20,000 (*P. graminis*, Table C.9) could indicate very large numbers of yet unknown genes or it could be indicative of gene prediction artefacts.
We identified 38 gene families which are much smaller in *B.g. tritici* and *P. graminis* than in the non-obligate biotrophs. We categorised them in three groups i) transporters ii) enzymes involved in redox reactions and iii) others (Figure C.1). Carbohydrate transporter families are small in mildew and rust, as was previously described for *P. graminis* (Duplessis *et al.*, 2011). *B.g. tritici* and *P. graminis* have only nine and 12 transporter genes, respectively (Figure C.1). In contrast, *A. nidulans*, *B. cinerea* and *M. grisea* have 55-96 genes encoding carbohydrate transporters. Several gene families involved in redox reactions are very large in these three but reduced in *B.g. tritici* and *P. graminis* (Figure C.1). These include proteins with FAD-binding domains, aldo-keto reductases, glucose-methanol-choline (GMC) oxidoreductases, flavin binding monooxygenases and the large family of genes that encode cytochrome P450 proteins. Loss of certain gene families involved in redox reactions were also found in the obligate biotroph rust of Arabidopsis (Kemen *et al.*, 2011; Raffaele and Kamoun, 2012). Gene families that are completely absent in powdery mildew and rust include enzymes that degrade the plant cell wall components such as cellulose, hemicellulose and pectin (CW, Figure C.2). Reduction in cell wall degrading enzymes has been previously reported in *Ustilago maydis* (Kämper *et al.*, 2006; Raffaele and Kamoun, 2012).
Recent studies found that biotrophic fungi lack certain enzymes that are involved in the sulfate and nitrate metabolism (Baxter *et al.*, 2010; Spanu *et al.*, 2010; Duplessis *et al.*, 2011; Kemen *et al.*, 2011). To investigate this topic in more detail, we did a genome-wide comparative analysis using enzymes of *Saccharomyces cerevisiae* biochemical pathways as a reference dataset. We found that mildews and rust fungi lack the exact same set of enzymes of the methionine/cysteine pathways (Figure C.3). In both, the three enzymes performing the transformation of anorganic sulfur (SO_4^{2-}) to sulfite (SO_3^{2-}) are missing. In addition, both lack three enzymes needed to produce siroheme, a prostethic group required for the function of sulfite reductase (MET10), the enzyme responsible to convert sulfite into sulfur hydroxide (H_2S, Figure C.3c). Curiously, both mildew and rust fungi still contain one subunit of sulfite reductase (MET10), which should not be functional without the prostethic group siroheme. It is therefore possible that MET10 performs a second function which does not depend on the prosthetic group.
Additionally, mildews and rusts have deficiencies in the synthesis and catabolism of tryptophan, phenylalanine and tyrosine, as they lack enzymes that perform the final steps of the biosynthesis as well as the first steps of degradation of phenylalanine and tyrosine. Again, mildews and rusts lack the exact same three enzymes (ARO8, ARO9 and ARO10) (Figure C.3a).

4.4.3 Comparative analysis of the *B.g. tritici* and *B.g. hordei* gene complements

All except 367 *B.g. tritici* genes have homologs in the *B.g. hordei* genome, indicating that the overall gene content of the two *formae speciales* is very similar. We identified 5,258 pairs of bi-

directional closest *B.g. tritici*/*B.g. hordei* homologs and aligned them both at the protein and at the DNA level (see Text 4.3.5). The CDS of the gene pairs are on average 93.7% identical. The predicted proteins are 93% identical and 95.8% similar. In total, 2,542,414 amino acid positions could be aligned of which 156,055 were polymorphic. The 367 *B.g. tritici* genes which have no homologs in barley powdery mildew also have no homology to other known proteins and also do not contain known protein domains (e.g. signal peptides).
Synonymous positions in the CDS alignments were used to estimate the divergence times gene pairs and overall species divergence (Text 4.3.6). The divergence time estimates for the 5,258 gene pairs is 6.3 $\pm$ 1.11 Myr. Thus, based on these CDS alignments, we estimate that *B.g. tritici* and *B.g. hordei* diverged between 5.1 and 7.4 Myr ago. This narrows down previous estimates which ranged from 4.7 (Inuma *et al.*, 2007) to 10 Myr (Oberhaensli *et al.*, 2011), indicating that the two *formae speciales* diverged several Myr ago, but clearly after the divergence of their hosts 9-15 Myr ago (Huang *et al.*, 2002; Akhunov *et al.*, 2003; Chalupska *et al.*, 2008).
Intergenic regions could only be compared to a limited degree. This is because intergenic sequences in both *B.g. tritici* and *B.g. hordei* contain very high numbers of repetitive sequence which are difficult to assemble and/or anchor. Nevertheless, we identified one large region of approximately 300 kb that allowed alignment of *B.g. tritici* and *B.g. hordei* genomic sequences. Out of 22 *B.g. tritici* genes, 18 genes were found in colinear positions, the few exceptions are likely due to mis-assemblies (Figure C.4). One *B.g. hordei* gene was completely absent from the *B.g. tritici* genome, indicating a gene loss in *B.g. tritici* (Figure C.4). Intergenic regions are strongly divergent due to differential insertions of TE sequence in the two *formae speciales*. These data agree with previous findings (Oberhaensli *et al.*, 2011).

4.4.4 Candidate effector proteins show characteristics of diversifying selection

For previous studies, the criterion for the identification of CSEPs was that the genes code for short proteins with a signal peptide but otherwise have no homology to known proteins (Godfrey *et al.*, 2010; Spanu *et al.*, 2010). We have identified 437 CSEPs in the *B.g. tritici* genome, all of them have homologs in *B.g. hordei*. To test whether CSEPs are under diversifying selection, we calculated for each of the *B.g. tritici*/*B.g. hordei* homologs the ratio of dS (synonymous substitutions per synonymous site) and dN (non-synonymous substitutions per non-synonymous site). We found that CSEPs have an average dN/dS ratio of 0.8. In contrast, non-CSEP genes gave an average dN/dS ratio of 0.24 and 96.6% of them have a dN/dS ratio of less than 0.5 (Figure 4.1). We therefore used the dN/dS ratio as an additional criterion to identify further effector gene candidates: gene pairs with an average dN/dS higher than 0.5 were considered as possibly being under diversifying selection. As a note of caution, we point out that we propose this method simply to identify possible effector candidates that do not match the common definition of small proteins with signal peptides. Since the proteins encoded by these genes do not have a signal peptide, we refer to them simply as candidate effector proteins (CEPs). These new CEPs are without exception gene families that have no yeast homologs. Thus, with the 437 CSEPs and the 165 novel CEPs, the *B.g. tritici* genome contains an arsenal of 602 putative effector genes.
To address the possibility that some CSEPs or CEPs might be pseudogenes, we analysed data

of the *B.g. tritici* post-infection transcriptome. We found that only four CSEPs and four CEPs showed no evidence of transcription. From the subset of 237 CSEPs which were used for the comparison with Bgh CSEPs 234 (99%) showed expression.

Candidate effector gene families

We pooled all CSEP and CEP encoding genes from *B.g. tritici* and *B.g. hordei* into a database in order to identify gene families. Interestingly the total number of CEPs and CSEPs in *B.g. hordei* is with 565 almost identical with that in *B.g. tritici*, although *B.g. hordei* has fewer CSEPs (420) and more CEPs (145). We grouped the 1,167 candidate effector proteins into 248 families. Most of them (162) have exactly two members, one from *B.g. tritici* and one from *B.g. hordei*. Only 8 gene families have more than 10 members (Table C.3). The largest family has 200 members and can be subdivided into several subfamilies. The family displays a series of highly conserved cysteine, tyrosine and phenylalanine residues. In general, cysteines are the amino acid residues which are the most conserved in families with more than 10 members. Most large families show differential expansion of subfamilies in either *B.g. tritici* or *B.g. hordei* (example in Figure C.5).
Because many CSEP genes are members of gene families, it is possible that differential deletions of homologs in the two *formae speciales* could lead to paralogs being wrongly identified as bi-directional closest homologs ("deep paralogs"). Deep paralogs should show a higher rate of synonymous substitutions than true orthologs because their actual divergence is further in the past and they had more time to accumulate substitutions. Indeed, Figure C.6a shows that dS rates in CSEPs are generally higher than in non-CSEPs (on average 1.4-fold). Consequently divergence time estimates for CSEPs are overall higher than for non-CSEPs (Figure C.6b). This indicates that a some *B.g. tritici/B.g. hordei* genes pairs are deep paralogs. This can also be illustrated with CSEP family FAM46-114 which contains four cases where the closest *B.g. tritici* and *B.g. hordei* homologs are placed in branches of the phylogenetic tree which diverged prior to the two *formae speciales* (Figure C.5). These deep paralogs also have estimated divergence times that predate the *B.g. tritici*/*B.g. hordei* divergence, while divergence times of putative orthologs mostly coincide with the estimated *B.g. tritici*/*B.g. hordei* divergence (Figure C.5).

4.4.5 The mating type locus of *B.g. tritici*

The mating type locus of *B.g. tritici* comprises either of the two mating type idiomorphs *MAT1-1* or *MAT1-2* and is flanked by the *SLA2* gene on one side and the *APN2/COX13* genes on the other. We have analysed the *MAT* loci of four *B.g. tritici* isolates, two isolates per idiomorph. In the reference sequence (isolate 96224), all mating type related genes (MAT idiomorph and flanking genes) were found on the same FP contig (Figure C.8). The genome sequences of the three isolates that were sequenced with Illumina (JIW2, 94202 and 70) are more fragmented. Therefore, the mating type genes are located on separate, rather short contigs.
Isolates 96224 and 70 have a *MAT1-2* idiomorph. In isolate 96224, the *MAT1-2-1* gene is located within 20 kb distance to the *SLA2* gene (Figure C.8). The locus itself is situated on a 454 sequence scaffold of 40kb size and the genes are surrounded by TEs from various families. *APN2* and *COX13* are located on the same FP contig at a distance of about 600kb. The *MAT1-2-1* gene of

isolate 70 was found on a 3.7 kb contig of the Illumina *de novo* assembly. The coding sequence of the *MAT1-2-1* gene of the two isolates is identical. It has 3 exons and encodes a protein with 343 amino acids that includes a HMG-box domain (aa 169-240). The *SLA2* genes of isolate 70 and isolate 96224 code for proteins of 1029 amino acid length which differ in two amino acids.
Isolates JIW2 and 94202 have a *MAT1-1* idiomorph which includes two genes, *MAT1-1-1* and *MAT1-1-3*. The *MAT1-1-3* and *MAT1-1-1* genes are located on different contigs of the Illumina *de novo* assemblies (Figure C.8). *MAT1-1-3* has 3 exons and encodes a protein of 326 amino acids, of which the HMG box is well conserved compared to the *E. necator* MAT1-1-3 (Brewer *et al.*, 2011). The MAT1-1-3 proteins of the two isolates differ in one amino acid. *MAT1-1-1* and *SLA2* are located on the same contig, *SLA2* upstream of *MAT1-1-1* as described for *E. necator*. The *MAT1-1-1* genes in JIW2 and 94202 are identical and encode a 241 amino acid protein. The alpha box of the protein shows similarity to the *E. necator* MAT1-1-1 protein, but the N- and C-terminal regions are not conserved. The SLA2 proteins of JIW2 and 94202 differ in one amino acid and are slightly larger than the proteins encoded by the opposite mating types (isolates 96224 and 70). The SLA2 proteins of the two mating types share 90% similarity. The *MAT1-1* idiomorph apparently does not contain any additional genes such as *MAT1-1-2* or *MAT1-1-4* which were found in other fungi (Debuchy and Turgeon, 2006).
The MAT1-2 proteins of *B.g. tritici* and *B.g. hordei* share 86% identity and 90% similarity, and the SLA2 proteins are 96.7% identical and 98.4% similar. In *B.g. hordei*, the *MAT1-2* and the *SLA2* gene were found in different loci (Spanu *et al.*, 2010). The *MAT1-2* idiomorph and the *SLA2* gene of *B.g. tritici* however are located in the same locus with only 20kb distance, and in the *MAT1-1* idiomorph the *SLA2* and *MAT1-1-1* genes are even adjacent. Furthermore, *COX13* and *APN2* are located on the same FP contig as the *MAT* idiomorph and the *SLA2* gene. In many ascomycete fungi, *APN2* and *SLA2* were identified as the flanking sequences of the mating type locus (Debuchy and Turgeon, 2006). This holds true as well for *B.g. tritici*, however *ANP2* and *SLA2* are located within a larger distance of 600kb.

B.g. tritici isolates of opposite mating types can be crossed easily

A cross between *B.g. tritici* isolates 96224 and JIW2 was first performed in 2007 at the John Innes Centre in Norwich (UK). Plants were grown in the greenhouse and the mildews growing on them produced sufficient chasmothecia to generate a segregating population. This cross was repeated at the University of Zurich (Switzerland) together with a new cross between isolates 96224 and 94202. Both crosses were done in greenhouses as well as in climate chambers. Under both conditions both crosses gave a high number of leaves bearing chasmothecia. Depending on the experiment, ejection of ascospores occurred between day 4 and day 12 after the chasmothecia were placed in high humidity conditions.

4.4.6 Gene loss in *B.g. tritici* isolates as result of DNA repair

We identified a total of 537 large deletions (>500 bp), 114 of them are shared between at least two isolates and only 15 are common to all three isolates. Most deletions affected intergenic regions which are mainly consisting of TEs. Only in 18 cases, genes are completely or partially deleted

in one or more of the three isolates (Table 4.1). Two of these presence/absence polymorphisms were genes of the mating type locus (see above). The other 16 genes were lost in deletions which span between 600 bp and 44 kb (Table 4.1). Only in two cases the gene was missing in all three isolates (Table 4.1). Interestingly, 13 of these 16 genes which are absent in some or all of the isolates are CSEPs or CEPs, one is a peptidyl isomerase and two are *ab initio* gene predictions without homologs in any other fungal genome.
To check whether there is a correlation between geographical origin of an isolates and presence/absence of a particular gene we studied the 16 genes in 6 additional *B.g. tritici* isolates, five Israeli and one Swiss isolate. As Table C.4 shows, there is no correlation between the presence or absence of a gene and the geographical origin of the isolate.
We wanted to study the possible molecular mechanisms that lead to the observed gene loss. In the case of gene *BgtE-5419*, the gene is part of a 8 kb deletion. Upon inspection of the deletion breakpoint we found a near-perfect 23 bp repeat motif that flanks the region of the deletion in 96224 (Figure 4.2a). This signature is a strong indication that this deletion is the result of the repair of a double-strand break (DSB) which occurred somewhere within the now deleted region. DSB repair via the single-strand annealing (SSA) pathway leads to deletions and typically leaves signatures as the ones observed (Agmon *et al.*, 2009).
One additional gene (*BgtE-5692*) is not actually absent in JIW2 and 94202, but differs strongly in sequence from the allele in 96224. While most of the gene and its surrounding sequence are virtually identical between the two alleles, a region of 140 bp has a much higher SNP density (Figure 4.2b). Possibly, the variable segment was introduced in a gene conversion event during DSB repair (Figure C.9). Gene conversion can happen between homologous chromosomes (homologous recombination) or between any homologous sequences in the genome and leads to transfer of genetic information from the intact locus to the region with DBS (Chen *et al.*, 2007).

4.4.7 Sequence diversity of four *B.g. tritici* isolates

We detected between 113,000 and 161,000 SNPs in the three isolates. The ratio between transitions and transversions is approximately 1.5 for all isolates. Insertion and deletions (InDels) of 1 to 4 bp (the size range that can be detected by mapping of Illumina reads) were almost 100 times less frequent than SNPs (Table C.5). The Swiss isolate 94202 and the UK isolate JIW2 show similar levels of polymorphism compared to the 96224 reference genome. With approximately 40% more SNPs and InDels, the Israeli isolate 70 is the most divergent of the three.
In all three isolates, between 3.7% and 3.9% of the SNPs were found in CDS of genes. Out of these, about 44-46% lead to amino acid changes in the predicted protein sequence (Table C.5). For each gene, the fraction of non-synonymous substitutions compared to the total number of substitutions was calculated and normalized by the length of the gene. Values were plotted separately for CSEPs/CEPs and all other genes (Figure C.10). CSEPs/CEPs contain on average more non-synonymous substitutions than all other genes.
Interestingly, the different isolates have genomic segments that clearly represent independent evolutionary lineages, as many genes were present in multiple protein variants. We defined protein variants as differences in the predicted amino acid sequences of a gene between different isolates. More than half of the genes (56.8%) show no variation in the predicted amino acid

sequence (i.e. the same protein variant is represented in the four isolates, Table C.6). 30% of the genes were present in two protein variants and 13% had three or four protein variants (Table C.6). Interestingly, CSEPs and CEPs tend to have more protein variants than non-CSEPs/CEPs. However, we did not test whether that difference is statistically significant.
To sample *B.g. tritici* more deeply, the sequences of 25 randomly selected CSEP and 25 non-CSEP genes were extracted from 6 additional isolates (see methods, S2.8), bringing the total of sampled isolates to 10. For each gene, we counted the total number of polymorphic sites and calculated the average number of polymorphic sites per 100 bp in multiple alignments of four (isolates 96224, JIW2, 94202 and 70) and all ten sequences. This was done to study whether the number of haplotypes increases with the number of isolates sampled. Alignments of CSEP genes with all 10 sequences had in average a 1.6 fold higher number of polymorphic sites per 100bp (0.445 polymorphic sites per 100bp) compared to alignments with only four sequences (0.273 polymorphic sites per 100bp). In non-CSEP genes, the difference was even higher with 1.7 fold higher average number of polymorphic sites per 100bp when 10 isolates where used (0.249 polymorphic sites per 100bp) as opposed to 4 isolates (0.145 polymorphic sites per 100bp). Again, we found no correlation between sequence similarity and geographical origin of the isolates.

4.4.8 McDonald-Kreitman like analysis of CSEP genes

To estimate putative adaptive substitutions in CSEPs, the 237 bi-directional *B.g. tritici/B.g. hordei* CSEP homologs (Figure 4.1) were analysed using a McDonald-Kreitman test (McDonald and Kreitman, 1991). In addition, the proportion of adaptive substitutions α and direction of selection (DoS) was calculated (Stoletzki and Eyre-Walker, 2011). For 12 CSEPs we did not have complete sequences for all four *B.g.tritici* isolates due to large sequence gaps or presence/absence polymorphisms. These were excluded from the test as well as 86 CSEPs that had no polymorphism between *B.g.tritici* isolates. Another 62 had to be excluded because they showed no synonymous substitutions (which would lead to divisions by zero), leaving a total of 77 CSEPs where the analysis could be performed.
It should be noted that there is relatively little statistical power with only four isolates. Additionally, there is sufficient divergence between *B.g. tritici* and *B.g. hordei*, but there is very little polymorphism among *B.g. tritici* isolates. In total, 7 genes showed a significant χ^2 value at $p < 0.05$ in the McDonald-Kreitman test and a positive direction of selection (DoS). Another 8 CSEPs closely miss statistical significance with χ^2 p values between 0.05 and 0.11, an $\alpha > 0.6$ and DoS > 0.3. In total, 55 genes had a positive direction of selection (DoS>0) and an $\alpha > 0.1$, but do not pass the McDonald-Kreitman test. A total of 47 CSEPs have an $\alpha > 0.5$. These 55 genes are good candidates of selection for future studies with larger numbers of isolates. It is common that only few genes pass the MK test. In that respect, the MK-like analysis provides strong evidence that the CSEPs analyzed are indeed under positive selection.

4.4.9 *B.g. tritici* isolates are mosaics of different haplogroups

We observed that the SNP frequency in all isolates varies strongly along the genome. In all three isolates, we found regions of variable size which are almost completely identical with the 96224

reference sequence. These regions are separated by breakpoints from regions with a several-fold higher SNP frequency (Figure 4.2c and 2d). These distinct segments of varying SNP frequencies indicate that the isolates studied are complex mosaics of different haplogroups. The more divergent haplogroup we referred to as H_{old} and to the one that is more similar to the 96224 reference sequence as H_{young}.
The regions of high or low SNP frequencies differ between the four isolates, indicating that they are the products of independent recombination events (examples in Figure 4.2c and 2d). Through pairwise comparison of the haplogroup profiles of the four isolates we identified between 366 and 618 genomic segments where isolates are of the same haplogroup (Table C.7). For example, in isolates JIW2 and 94202, we identified 622 genomic fragments (24.2% of the analysed sequence) that belong to the same haplogroup H_{young}. Interestingly, the Israeli isolate 70 has in all comparisons the lowest fraction of H_{young} segments as only between 6.9 and 8.6% of its genome is of the same haplogroup as any of the other isolates (Table C.7).
Only very few regions (approximately 2.2% of the genome) are of the same haplogroup H_{young} in all four isolates. We examined these regions in more detail because they could represent loci indicating selective sweeps. However, we did not find particular types of genes in these regions. In fact, some of the invariable regions contained no genes at all.
We identified 518 genomic segments (approximately 24.5% of the genome) where the three European isolates 96224, 94202 and JIW2 all differ from each other, indicating that these regions are derived from three different ancient haplotypes. Inclusion of the Israeli isolate 70 in the comparison led to the identification of 471 genomic fragments (approximately 17.4% of the genome) where even four different haplotypes can be distinguished.

4.4.10 Haplogroup segment numbers suggest frequent asexual propagation and inbreeding

We identified between 764 (isolate 70) and 1312 (isolate JIW2) haplogroup breakpoints across the genomes (see methods). Thus, the average size of the haplogroup segments is similar in all three re-sequenced isolates, ranging from 87.3 kb (isolate JIW2) to 150 kb (isolate 70). Considering the possibility that some sequence scaffolds might not be anchored in the correct order, the actual length of haplogroup segments in the *B.g. tritici* genome is probably somewhat longer. To estimate the number of sexual recombinations that are necessary to obtain the observed number of haplogroup breakpoints, we simulated recombination in haploid populations. For the simulations we assumed that the genetic map of *B.g. tritici* has as size of approximately 2,000 cM (i.e. 20 recombinations per progeny individual). That is based on the genetic map of *B.g. hordei* which has a size of 2,114 cM (Pedersen *et al.*, 2002a). The simulations were run in two ways:
First, we assumed that two large populations (at least 1,000 individuals) representing two different haplogroups can mate without restrictions. Under these conditions, the average number of haplogroup breakpoints increases in a linear manner. For a genetic map of 2,000 cM, each generation has on average 10 breakpoints more than the previous one (Figure C.13a). Thus, the observed numbers would be obtained in 76 (for isolate 70) to 131 sexual generations (for isolate JIW2). Second, we studied the influence of population size on the number of haplogroup breakpoints over long periods of time (e.g. 1,000 sexual generations). In small populations,

the number of haplogroup breakpoints stabilises relatively quickly after an initial increase. For example, a population of size 40 will converge towards genomes with approximately 600 haplogroup breakpoints within approximately 200 sexual generations (Figure C.13b). Thus, smaller populations will converge toward fewer haplogroup breakpoints than larger ones due to this inbreeding effect. In very small populations (e.g. 8 individuals), the genotypes reach a state analogous to homozygosity in less than 50 generations (Figure C.13a). At this stage, the individuals differ only in their mating type loci while the rest of their genomes are identical. Our simulations show that the number of breakpoints converge close to the observed values at small population sizes of 40-100 (Figure C.13b).
From these data we conclude that the observed patterns can either be explained by inbreeding of a very small population or through a small number of sexual generations in a large population. Either explanation indicates that the isolates studied propagate through asexual (clonal) reproduction and/or near-clonal reproduction resulting from inbreeding in small populations.

4.4.11 Comparison of sequence diversity in *B.g. tritici* isolates and in *Phytophthora* species

We compared the level of sequence diversity between *B.g. tritici* isolates with those previously published for *Phytophthora* species (Raffaele *et al.*, 2010). The three distinct species *P. ipomoeae, P. mirabilis* and *P. phaseoli* differ from the reference genome of *P. infestans* in 438,804 to 652,688 SNPs (Raffaele *et al.*, 2010). The 240 Mbp *Phytophthora* genome is similar in size to that of *B.g. tritici*. Therefore, the overall levels of sequence divergence between *Phytophthora* species are roughly 3-5 times higher than between *B.g. tritici* isolates. In contrast, the two *formae speciales* wheat and barley powdery mildew (which diverged about 6.2 Myra go) are so highly divergent that an assessment of SNP numbers is no longer possible. These data indicate that at the sequence level, *Phytophthora* species are very closely related and probably only diverged a few hundred thousand years ago (assuming a similar mutation rate for *Phytophthora* and *Blumeria*).

4.5 Supplementary text: discussion

4.5.1 Specific gene loss in biotrophs reflects their life style

We compared the gene complements of obligate biotrophic fungi *Blumeria* and *Puccinia* with those of oligotrophs (*A. nidulans*) and hemibiotrophs (*M. grisea* and *B. cinerea*). Interestingly, the obligate biotrophs show a strikingly similar pattern of gene loss, despite their vast phylogenetic distance. The largest differences in sizes of gene families were found in enzymes involved in redox processes. A previous study suggested a drastic reduction in the secondary metabolism in obligate biotrophs as key enzymes such as polyketide synthases (PKSs), terpene cyclases, and di-methylallyl diphosphate tryptophan synthases are missing in obligate biotrophs (Spanu *et al.*, 2010). Our finding of the drastic reduction of enzymes involved in redox reactions supports this hypothesis. Redox enzymes are frequently involved in secondary metabolic pathways such as toxin production. For example 10 out of 17 enzymes in the aflatoxin synthesis pathways perform redox reactions (Yu *et al.*, 2004).

Ours as well as previous studies (Kämper *et al.*, 2006) showed that biotrophs also lack many cell-wall degrading enzymes, *Blumeria* to a greater extent than rust. This indicates that cell wall penetration is based specifically on mechanical force provided by the appressorium and no enzymatic degradation of the cell wall is involved. In contrast, hemibiotrophs or autotrophs need many such genes because the degradation of plant cell wall material is a major source of carbohydrates. It is also possible that the lack of hydrolases reflects a selective pressure, as such enzymes or their generated products may be recognised more easily by the plant (Dong *et al.*, 2011).

As previously described (Spanu *et al.*, 2010; Duplessis *et al.*, 2011; Kemen *et al.*, 2011), both mildews and rusts lack the exact same set of enzymes of the methionine/cysteine biosynthetic pathways. The network of the sulfur assimilation pathways that leads to the synthesis of sulfur-containing amino acids has been reduced dramatically and identically in both. Since ascomycetes and basidiomycetes have diverged long before the mildews and yeast, these gene sets must have been eliminated independently in both evolutionary lineages. Additionally, we identified identical gene loss in the phenylalanine and tyrosine metabolic pathways, similar to what was described in *B.g. hordei* (Spanu *et al.*, 2010). This indicates that the evolution to obligate biotrophy follows very distinct paths that lead to identical outcomes (Spanu, 2012).

4.5.2 Does selection pressure drive specific gene loss in *B.g. tritici*?

We identified hundreds of large deletions in the genomes of the three isolates. Because most are in intergenic (i.e. TE) sequences, they most likely have little effect on the fitness of the organism. Only 16 large deletions led to gene loss. We propose that repair of double-strand breaks (DSBs) as was previously described for plants (Agmon *et al.*, 2009; Wicker *et al.*, 2010; Buchmann *et al.*, 2012) led to the deletion of these genes. DSBs probably occur largely randomly across the genome. Repair of these breaks consequently leads to randomly distributed deletions. Deletions must occur frequently as is reflected in the hundreds of deletions in TE sequences. Thus, genes must be affected by deletions also on a random basis. Most gene deletions are probably delete-

rious and are therefore selected against.
It is intriguing that 13 out of 16 gene deletions affected genes encoding putative effector genes. Considering that CEPs and CSEPs make only about 10% of the gene content, they are tremendously over-represented in these presence/absence polymorphisms (two-tailed Fishers exact test: p=1.3E-11). Additionally, two of the deleted non-CEP genes were mere predictions without homologs in other fungal species. Thus it is possible that these are gene prediction artefacts, which would make the bias toward CEP/CSEPs genes even greater. Simple redundancy in CEP/CSEP gene families does not explain this bias. It was shown in fission yeast that large numbers of genes can be deleted without a visible phenotype and only 17% of all genes were found to be essential for viability (Kim *et al.*, 2010). Thus, our data suggest specific selection pressure for loss of these putative effector genes. However, selective loss is only one possible explanation. Fungal effectors are often functionally redundant. It is therefore possible that some of the gene losses are simply the result of genetic drift.
Comparison of orthologous loci from *B.g. tritici* and *B.g. hordei* showed that intergenic sequences are not conserved between the two *formae speciales*. This "genomic turnover" is driven by the amplification of TEs and the removal of DNA through random deletions (Oberhaensli *et al.*, 2011), a process that has also been described for plant genomes (Wicker *et al.*, 2003; Vitte and Panaud, 2005). Thus, deletions must from time to time also affect genes. If a deletion removes an effector gene whose product was recognised by the host plant, this deletion will become rapidly fixed in the gene pool as it increases the fitness of the pathogen. However, this comes at the price of a slow but steady draining of the reservoir of effector genes in the *B.g. tritici* gene pool. Indeed, compared to human or animal pathogens, selection pressure on a crop plant pathogen must be disproportionately stronger because of intense crop breeding for new resistant varieties. In contrast, selective breeding for pathogen resistance is a lot less frequent in animals and non-existent in humans.

4.5.3 Inbreeding and clonal propagation are important factors in *B.g. tritici* evolution

It is intriguing how much the genomes of the four powdery mildew isolates differ from each other. Until now, powdery mildew isolates were distinguished by their virulence to specific host varieties. Such virulence or avirulence can, in theory, be caused by one single point mutation that alters or destroys an effector gene. However, we found that isolates differ in the sequences of thousands of genes, many of them are CEPs and CSEPs. Even the two Swiss isolates differ as much from each other as they differ from the British isolate. Furthermore, the genomes are mosaics of very distinct haplogroups, some of which diverged 55,000 to 65,000 years ago.
These findings raise the question how these distinct genotypes evolved. If different lineages and haplogroups had been mixed continuously since tens of thousand of years, haplogroups would be completely homogenised and haplogroup segments would not be recognisable anymore. Our simulations suggest two explanations for the observed numbers of haplogroup breakpoints: First, it is possible that a large population (e.g. 1,000 individuals) population originally consisted of two sub-populations of different haplotypes which were able to mate without restrictions. Such unrestricted mating between all individuals of the population would result in

genomes with the observed haplogroup patterns after approximately 100 sexual generations.
The second explanation is that small populations propagate through inbreeding over long periods of time. In this case, the number of haplotype breakpoints would stabilise at a certain level because the inbred population has reached a state of "homozygosity" (i.e. individuals only differ in their mating type locus while their genomes are otherwise identical). At this point, sexual propagation would be largely clonal (except for the inevitable difference in the mating type loci). Here, it is necessary to point out that in the haploid *B.g. tritici* system, "homozygosity" refers to the origin of haplotype segments in a genotype (synonymous with autozygosity or "identical by descent").
Clonal reproduction has been shown to be extremely important for fungal pathogens in animals and humans as it preserves successful combinations of genes and alleles (e.g. *Candida albicans* and *Cryptococcus neoformans*, (Heitman, 2006; Bougnoux *et al.*, 2008; Tibayrenc and Ayala, 2012)). In contrast, sexual recombination can lead to the loss of effective virulence factors or to the acquisition of undesirable avirulence genes (Heitman, 2006; Bougnoux *et al.*, 2008). Nevertheless, rare sexual or parasexual recombination is viewed as essential for successful adaptation to changes in environment or for drug resistance (Bougnoux *et al.*, 2008; Heitman, 2006).
In *Blumeria*, both the asexual and the sexual cycle regularly occur. Thus, individual powdery mildew isolates could propagate purely clonally through asexual spores that over-winter because of their vast numbers and the availability of "green bridges" (Liu *et al.*, 2012). Alternatively, successful genotypes could be maintained through inbreeding. Our simulations show that inbreeding in populations of 40-100 individuals leads to homozygosity within 200 to 500 generations (Figure C.13).
In summary, we propose that *B.g. tritici* isolates mostly propagate in a clonal or near-clonal manner (which may include some inbreeding) (Tibayrenc and Ayala, 2012). The trigger for the formation of a new mildew isolate is the recombination of different haplogroups. We speculate that specific gene combinations which are contributed by the two different haplogroups determine the virulence (i.e. success) of the new isolate. These precise combinations of genes are strongly selected for, thus allowing only asexual propagation or a very limited sexual recombination between individuals that contain the successful gene combinations. The sexual recombination between the successful genotypes leads to rapid inbreeding. This highly inbreed population can then propagate in a quasi-clonal manner while the presence of both mating types still allows the regular production of winter-hard spores. Therefore, the potential of recombination by outbreeding with other mildew isolates and thus the possibility to adapt to changing environments such as host plants with new resistance genes would be maintained.
This situation is reminiscent of that in *Magnaporthe oryzae* where asexual propagation is predominant in most parts of the world. However, in *Magnaporthe oryzae* sexual recombination seems to be restricted to the center of origin of the pathogen (the Himalayan foothills) while most strains that have spread asexually all over the world have lost female fertility (Saleh *et al.*, 2012a,b). This is in contrast to *Blumeria* where isolates obviously have maintained their ability to reproduce sexually (as our crossing experiments show) despite a predominantly asexual propagation. In this respect *Blumeria* also differs from the hemibiotroph *Colletotrichum* species which seem to propagate exclusively asexually (Dean *et al.*, 2012).

4.5.4 *B.g. tritici* evolution differs from that of hemibiotrophic pathogens

Most wheat varieties used nowadays in agriculture are hexaploid (*Triticum aestivum*) which originated only approximately 9,000-12,000 years ago (Salamini *et al.*, 2002). However, the genomes of the *B.g. tritici* isolates (which all have hexaploid wheat as host) are composed of haplogroups which diverged already at least 48,000 years ago. Thus, the initial radiation of *B.g. tritici* haplogroups must have occurred on tetraploid wheats as hosts which existed for at least 300,000 years (Dvorak and Akhunov, 2005; Haudry *et al.*, 2007). We conclude that the host shift from wild tetraploid to domesticated hexaploid wheat did not result in a reduced genetic diversity in *B.g. tritici*.

It is known that *B.g. tritici* isolates can infect tetraploid as well as hexaploid wheat, but there appears to be some degree of "fine tuning" with specific isolates preferring hexaploid or tetraploid wheat (Eshed *et al.*, 1994). This suggests that the *B.g. tritici* gene pool provides all necessary genetic diversity for adaption to a range of tetraploid and hexaploid wheat species. Obviously, our sample of four completely sequenced isolates can not sufficiently assess the diversity of the *B.g. tritici* gene pool. However, a previous analysis of 12 genes from 203 *B.g. tritici* isolates (141 from the US, 34 from UK and 28 from Israel) (Parks *et al.*, 2009) identified the presence of 25 different haplotypes. Diversity was highest in UK and Israeli isolates where 14 different haplotypes were identified. Additionally, our own additional sampling of isolates showed that the number of polymorphic sites of 50 randomly selected genes increased approximately 1.6 fold when sequences from 6 additional isolates were added to the four isolates presented in this study (i.e. in total 10 sequences per gene). These data support our conclusion that a large diversity of *B.g. tritici* races can attack hexaploid wheat.

This host range expansion is strikingly different from *Phytophthora* and *Mycosphaerella* species where host changes went along with the formation of new species and loss of genetic diversity (i.e. bottlenecks, (Haas *et al.*, 2009; Stukenbrock *et al.*, 2010, 2012)). In the case of *M. graminicola*, the occurrence of the new host *Triticum aestivum* is even believed to be the trigger for the formation of the species (Stukenbrock *et al.*, 2011).

Millions of years of evolution have not led to species formation in *Blumeria*. In contrast, the genera *Phytophthora* and *Mycosphaerella* consist of different, albeit closely related species. In fact, *M. graminicola* arose as a new species only about 11,000 years ago (Stukenbrock *et al.*, 2010) while the hybridisation of two *M. graminicola* haplotypes that gave rise to Z. pseudotritici probably dates back merely 500 years (Stukenbrock *et al.*, 2012). It is unclear what caused the reproductive isolation of these very closely related species, but it was suggested that a series of specific changes in mating type genes are at least in part responsible for the formation of the species barrier (Stukenbrock *et al.*, 2011). This situation is very different from what we find in *B.g. tritici*. The genomes of the *B.g. tritici* isolates are composed of segments from a much more ancient and divergent gene pool than for example those in *M. graminicola* and *Z. pseudotritici*. And yet, our own experiments have shown that they can be crossed without difficulties (Text 4.4.5). There is not even a species boundary between *B.g. tritici* and *B.g. hordei* formae speciales because mating and production of fertile offspring between them is possible (Hiura, 1978), although we found that they are at least 80 to 100 times more divergent at the DNA level than the various species of the *Phytophthora* and *Mycosphaerella* genera.

The lack of species boundaries between *Blumeria formae speciales* and the presence of ancient

haplogroups in the genomes of *B.g. tritici* isolates is in strong contrast to the rapid and frequent formation of new species in *Phytophthora* and *Mycosphaerella*. Furthermore, the possible absence of genetic bottlenecks suggests a greater ability for adaptation to new host species in powdery mildew than in the hemibiotrophs *Phytophthora* and *Mycosphaerella*.

The situation in *Blumeria* is somewhat similar to that in the hemibiotroph *M. oryzae* where different isolates can attack distantly related grasses such as rice or wheat (Couch and Kohn, 2002). These isolates clearly belong to the same species as they can be crossed and avirulence/virulence segregates in a 7:1 ratio, indicating that it is determined by only three genes (Tosa *et al.*, 2006). This simple heritability pattern suggests a very close phylogenetic relationship of the two isolates. However, an important difference is that *B.g. tritici* isolates that attack hexaploid bread wheat still largely maintain their pathogenicity on tetraploid wheat species (Eshed *et al.*, 1994) while in the described *M. oryzae* isolates, specific gene combinations make the fungus incompatible with the host. In this respect, the situation is more similar to that in the hemibiotroph fungi of the *Colletotrichum* genus where individual strains can attack multiple host species (Dean *et al.*, 2012).

CHAPTER 5

General discussion

5.1 Quality of the *B.g. tritici* reference sequence

De novo sequencing of a genome requires much more effort than re-sequencing of a known genome. The complexity of a *de novo* sequencing project is determined by the genome size and the percentage of repetitive DNA. Many fungal genomes are small (less than 40Mb) and not very repetitive, which makes them suitable candidates for sequencing using NGS technologies. *Blumeria graminis* genomes however are exceptionally large (*B.g. hordei* 120 Mb Spanu *et al.* (2010), *B.g. tritici* 180Mb) and have a very high content of repetitive elements (*B.g. tritici* about 90% TEs). Recently published fungal genomes with similar characteristics, i.e. *Tuber melanosporum*, *Melampsora populina*, *Puccinia graminis*, as well as *B.g. hordei* were sequenced by Sanger sequencing (Table 5.1, Martin *et al.* (2010a); Duplessis *et al.* (2011); Spanu *et al.* (2010)). Why it was possible to sequence the *B.g. tritici* genome with NGS with a resulting quality that is comparable to genomes sequenced by Sanger technology only (Table 5.1), is discussed in the following paragraphs:

Genomic DNA of isolate 96224 was sequenced with Roche/454 technology which at that time generated reads of about 400bp average length. The use of a middle size read length, paired-end data with a 3kb insert size and the integration of 20'000 BAC end sequences (Sanger, 600bp average length) led to a *de novo* assembly of only 3,522 scaffolds (assembly size: 97.4 Mb including sequence gaps). In comparison, a *de novo* assembly of the same isolate based on Illumina data only resulted in over 96,000 contigs (assembly size: 66.7 Mb).

By anchoring the 3,522 scaffolds from the 454 assembly to a BAC library Finger Print (FP) assembly (Parlange *et al.*, 2011) the number of "pseudomolecules" was reduced to 250. This is excellent compared to other genomes of comparable size and repetitiveness (*T. melanosporum*: 398 scaffolds, *M. populina* 462 scaffolds, *B.g. hordei*: 6,898 scaffolds, Table 5.1) and makes studies on large-scale genome structure feasible.

Out of 6,540 *B.g. tritici* gene models in total, 5,398 (82%) were predicted based on homology to manually curated *B.g. hordei* genes (Spanu *et al.*, 2010). The high quality of the reference gene set (*B.g. hordei*) makes our predictions more solid than pure *ab initio* prediction and reduces the chance for prediction artefacts.

The high abundance of TEs increases the complexity of the *Blumeria graminis* genomes consider-

ably. TEs or TE-related sequences that are mis-annotated as genes must be avoided: The more than 1,000 *B.g. hordei* EKA homologs which were annotated in *B.g. hordei* (Sacristán *et al.*, 2009; Spanu *et al.*, 2010) immensely complicated gene prediction in *B.g. tritici*. Preliminary work on low coverage assemblies and BAC sequences (described in chapter 2 and 3) contributed to our knowledge about the TE population in *B.g. tritici* and *B.g. hordei* genomes. The repeat libraries that were generated - a high quality repeat library, including protein sequences (chapter 2), and a low-quality but very exhaustive TE library (REPET, chapter 4) - became very useful during the gene prediction process. These libraries will facilitate future studies in *Blumeria graminis* and might be useful to distinguish between real genes and TEs that are transcribed. In several fungal genomes, effectors are located mostly in TE-rich regions (e.g. *P. infestans*). *Leptosphaeria maculans* and *M. grisea* are only two examples where gain of function due to transposon insertion, deletion or other genomic rearrangements have been observed (Stergiopoulos and de Wit, 2009). Thus, it is important to know as much as possible about the repetitive fraction of a fungal genome.

Quality control of large-scale datasets, such as *de novo* sequencing projects, is difficult. Many bioinformatic processes are subject to a trade-off between accuracy and time investment. Quality standards are achieved via a "plan-do-check-improve" approach, thereby testing a certain number of random samples repeatedly. This affects *de novo* or reference assembly accuracy, semi-automated processes such as repeat masking or SNP discovery, and most importantly gene prediction. The use of the reference sequence for further studies, e.g. for the development of molecular tools (see 5.3), will help to further asses it's quality (e.g. accuracy of assembly, gene models and SNP maps) and reveal areas for improvement.

5.2 How to improve the *B.g.tritici* sequence and gene annotation

Genome sequences, especially those of model species, are subject to constant revision as technology advances and knowledge improves (e.g. the *Saccharomyces cerevisiae* genome published in 1996 is today at release 64). The following paragraphs discuss how the quality of the *B.g. tritici* reference sequence could be improved:

The finger print assembly determines the entities (FP contigs) which provide the backbone of the genome sequence. The use of the LTC software (Frenkel *et al.*, 2010) instead of the FTP could possibly reduce the number of FP contigs and improve the accuracy of the contig assembly.

The large number of TE copies complicated the *de novo* assembly process, resulting in a fragmented assembly (3'500 scaffolds) and numerous sequencing gaps within scaffolds. Additional paired-end sequences with a large insert size that span the size of TE copies (approx. 10kb) could substantially improve the scaffold N50 of the assembly. In fact, 10kb paired-end reads were initially planned to be included. Unfortunately, this was technically not feasible due fragmented DNA: The biotrophic lifestyle of *B.g. tritici* makes it difficult to extract good, high-molecular weight DNA in large amounts which is needed for large insert libraries. Hopefully, technical advances in library preparation will lead to lower DNA requirement.

Third-generation sequencing technology (Pacific Biosciences, IonTorrent) which offer longer reads might also be an option that could lead to longer scaffolds. However, for a complex genome such as *B.g. tritici* which harbors thousands of almost identical TE copies, a high sequence accuracy is crucial. Therefore, it is very important to only use high accuracy sequence reads or reads that were corrected for sequencing-technology specific errors (introduced e.g. from PacBio technology).

As there were no *B.g. tritici* expression data available, the *ab initio* gene prediction software was trained on *B.g. tritici* gene models which were predicted based on homology to *B.g. hordei* genes. Surprisingly, out of totally 1'500 *ab initio* genes only 400 had no homolog in other fungal genomes suggesting a low rate of possible prediction artefacts. Still, extended quality control especially on the *ab initio* predicted gene models would be beneficial, for example through RNA seq data which could also provide additional information about untranslated regions of the coding sequences (UTRs). Confirmation of the *ab initio* predictions by a second *ab initio* gene prediction software might also be advisable, which was not feasible in the given timeframe of the project.

Ab initio gene prediction was significantly improved by running the prediction software on repeatmasked sequences. Comprehensive repeat libraries that include nucleotide as well as protein sequences of TEs are a fundamental requirement for accurate repeat masking. In this work, we combined a high quality TE library (chapter 2) with a "quick-and-dirty" REPET consensus library for the masking of the genome (Flutre *et al.*, 2011). Currently, there is redundancy between a large number of consensus sequences in the REPET library. In addition, many TE families of both libraries are not well characterized or even remain unclassified, among them some of the most abundant TEs in the *B.g. tritici* genome. The better the quality of the libraries is, the more accurate the masking will be. Accurate masking will lower the chances that non-repetitive sequences hidden in TE regions are masked (e.g. effectors). Efforts to reduce redundancy of elements in the REPET library and to improve characterization of "unclassified" and "unknown" TEs are therefore needed.

About 1,615 454 scaffolds (15Mb) could not be anchored to the FP backbone because no corresponding BAC end sequence was available. These orphan scaffolds were randomly arranged into a large artificial contig with a total size 19Mb (incl. gaps). Obviously, these un-anchored scaffolds are missing in the 250 FP contigs, which results in a number of large sequence gaps. Additional anchoring of the remaining scaffolds would be highly beneficial since 820 genes are located on them.

5.3 Benefits of a *B.g. tritici* reference sequence for powdery mildew research

The *B.g. tritici* reference sequence can be used as a tool to facilitate genetic mapping, for example for marker development. SNP maps that were generated for three *B.g. tritici* isolates compared to the reference (described in chapter 4) have already been used to develop KASPar markers. These markers can for example be used for gene mapping (e.g. *B.g. tritici* avirulence genes) or

to study population structure in *Blumeria graminis*. Other approaches such as mapping by sequencing (bulk sequencing) and re-sequencing of UV-mutants are as well based on the principle of variant detection compared to a reference sequence.
SNP detection in a complex genome such as *B.g. tritici* has to be done with caution. First, one has to be aware of potential inaccuracies in the reference sequence due to sequencing errors. Second, the high abundance of identical TE copies requires very stringent mapping parameters in order to reduce false positive SNPs. In *B.g. tritici*, high-confidence SNPs are mainly found in gene-encoding sequences, and therefore marker development should focus on these regions. In addition, it is important to validate *in silico* generated data whenever possible by additional methods such as PCR and Sanger-sequencing on genomic DNA. This is especially important when working in regions with a high amount of repetitive DNA or TEs because assembly and mapping algorithms may have problems to handle these correctly.
The re-sequencing of three additional isolates has provided a first insight into genome dynamics of *B.g. tritici* and genotypic diversity among isolates (chapter 4). These data allowed to speculate on the basis of virulence and avirulence of *B.g. tritici* isolates (e.g. CSEP/CEP genes which are under positive selection or presence/absence polymorphisms) and on the evolution of *B.g. tritici*. Additional sequence data of *B.g. tritici* isolates from diverse geographical regions and from different hosts (e.g. hexaploid and tetraploid wheat) would strengthen the conclusions drawn based on data of four isolates (in progress) and might most likely provide new findings which were not visible in a sample size of four. The genome sequences of two additional *B.g. hordei* isolates which are about to become publicly available (Hacquard *et al.*, submitted) will allow comparative approaches which are statistically more powerful.
Specificity of *Blumeria graminis* for cereal hosts has been a subject of interest for many researchers since decades. Based on numerous infection experiments, theoretical models were developed to explain the evolution of *Blumeria graminis* host specificity (Tosa, 1992). Now that two complete genomes (*B.g. tritici* and *B.g. hordei*, Spanu *et al.* (2010)) and a fragmented one (*E.pisi*, Spanu *et al.* (2010)) are available, observations made based on classical genetics can be combined with genomics data. Ideally, powdery mildew of rye (*Blumeria graminis* f. sp. *secalis*) and oat (*Blumeria graminis* f. sp. *avenae*) would be included as well. However, it is in any case important to conduct two-way comparative analyses. An analysis where e.g. only *B.g. tritici* genes are used which originate from homology-based gene prediction based on *B.g. hordei* is problematic. *B.g. tritici* genes which were predicted *ab initio* and are not present in *B.g. hordei* could be *forma specialis* specific genes and therefore candidates for host specificity determinants. It is important to asses them more in detail and to make sure that they are not prediction artefacts nor mis-annotated TEs.
The *in silico* prediction of effector candidates is a widely used application of genome sequences of fungal pathogens or symbionts. In the case of fungal and oomycete effectors, it is assumed that they have a signal peptide for secretion, particular sequence motives such as RXLR (oomycete, Whisson *et al.* (2007)) or YxW (*Blumeria graminis*, Godfrey *et al.* (2010)) but no homology to other known genes. The computational identification of effector candidates can be done relatively easily in an automated way and usually leads to several hundred candidate sequences. However, until a function involved in infection or resistance triggering can be proven, these sequences remain simply 'candidate effectors'.

5.4 Next-generation sequencing technologies: benefits and challenges for *de novo* sequencing of fungal genomes

The emergence of NGS technologies has boosted *de novo* genome sequencing of fungi. Nonetheless, most of the recently published high quality genomes of fungal pathogens or symbiont genomes have been sequenced with Sanger technology using BACs, plasmids and fosmids (Table 5.1). *De novo* assemblies of genomes larger than 40Mb with a high content of repetitive DNA which were sequenced using only short reads are usually quite fragmented and, therefore, consist of a large number of small contigs: The *Albugo laibachii* genome (size 37 Mb) which was Illumina sequenced has in its present state more than 3,700 contigs, and *Hyaloperonospora arabidopsidis* sequenced with a combination of Illumina and Sanger has more than 1,700 contigs (Table 5.1). The 5,000 contigs of the *de novo* assembly of the *Sordaria macrospora* genome (40 Mb, combination of Illumina/454) could only be placed into 152 scaffolds by comparison with reference genomes of three closely related *Neurospora* species (Nowrousian *et al.*, 2010). Fragmented assemblies are not a problem for studies on genetic diversity because the gene space is usually completely present if there is sufficient sequencing coverage. However, based on such fragmented genomes it is not possible to asses genome features at the chromosome level. This is unfortunate as high quality genomes can reveal highly interesting biological features: In *Fusarium* for example, massive chromosomal rearrangements and genes or chromosomal fragments acquired by horizontal gene transfer were found to be linked with pathogenicity (Ma *et al.*, 2010).

NGS technologies are highly suitable for *de novo* sequencing projects such as bacterial genomes, but obviously there is a limitation for larger genomes, particularly repetitive ones. Pushed by the pressure of publishing new findings with NGS, more and more low quality genome sequences become available, which can make it difficult to receive funding for a high-quality genome sequence (since "it is already published"). Sequencing projects with the aim of producing a high-quality genome sequence are time-intensive, laborious and expensive. However, extra effort eventually pays off because low-quality annotation or assemblies only allow low-level analyses (also known as the GIGO principle: garbage in, garbage out). Sequencing platforms like PacBio and Roche 454 have realized that there is a need for longer reads. However, significant improvement has to be made regarding error rate (PacBio) and cost efficiency (454).
The fungal kingdom is estimated to comprise about 1.5 million species (Hawksworth, 1991). Many of them have an enormous impact on the ecosystem as decomposers, symbionts or pathogens. The post-genomic era has begun for only very few fungal species. The 1'000 fungal genome project (F1000, funded by JGI in 2011) was started with the goal to "provide broad genomic coverage of the Kingdom Fungi", which will beyond doubt substantially contribute to our understanding of fungal biology. However, key to accurate analysis of these data are again well-characterized, foundational reference genomes [1].

[1] `http://1000.fungalgenomes.org/home/`

Table 5.1. A selection of sequenced (and not yet sequenced) fungal plant pathogens and symbionts

Fungus	Class.[a]	Lifestyle	Genome size Mb	Method[b]	Quality[c]	Genes	Additional information	Reference
Biotrophs and hemi-biotrophs								
Ustilago maydis	b	biotroph	21	WGS (p,f), 10x	24 sc, 23 chr	6,900	mating required for infection	Kämper *et al.* (2006)
Magnaporthe oryzae	a	hemibiotroph	40	WGS, 7x	159 sc	12,000		Dean *et al.* (2005)
Fusarium graminearum	a	hemibiotroph	36	WGS (p,f,BAC)	43 sc	11,600	4 chr, high/low SNP regions	Cuomo *et al.* (2007)
Laccaria bicolor	b	symbiont	65	WGS (p,f)	665 sc	20,000		Martin *et al.* (2008)
B.g. hordei	a	biotroph	120	WGS (p,f), solid	6,898 sc	5,854		Spanu *et al.* (2010)
Tuber melanosporum	a	symbiont	125	WGS (p)	398 sc	7,496		Martin *et al.* (2010a)
Mycospharella graminicola	a	hemibiotroph	40	WGS, 8x	21 chr	11,000	dispensome (HGT)	Goodwin *et al.* (2011)
Melampsora larici-populina	b	biotroph	101	WGS (p,f), 7x	462 sc	16,399		Duplessis *et al.* (2011)
Puccinia graminis triticina	b	biotroph	89	WGS (p,f)	392 sc	17,700	dicaryotic	Duplessis *et al.* (2011)
Leptosphaeria maculans	a	hemibiotroph	45	WGS (p+f)	76 sc	12,469	AT-rich regions with effectors	Rouxel *et al.* (2011)
Puccinia striiformis tritici	b	biotroph	78	Illumina	29,178 ctg	20,423		Cantu *et al.* (2011)
Hyaloperonospora arabidopsidis	o	biotroph	100	Sanger/Illumina	1,783 (>2kb)	14,500		Baxter *et al.* (2010)
Albugo laibachii	o	biotroph	37	Illumina 400/800PE	>3,700 ctg	13,000		Kemen *et al.* (2011)
Albugo candida	o	biotroph	45	454 PE (7kb), 20x	252 sc	15,800		Links *et al.* (2011)
Colletorichum graminicola	a	hemibiotroph	50	454/Sanger	13 chr	12,006		O'Connell *et al.* (2012)
Melampsora linii	b	biotroph	unknown	not sequenced	-	-	several AVRs cloned	Dodds *et al.* (2004, 2006)
Necro- and saprotrophs								
Stagonospora nodorum	a	necrotroph	37	WGS (p,f), 10x	107 sc	10,700		Hane *et al.* (2007)
Sclerotinia sclerotiorum	a	necrotroph	38	Sanger	36 sc, 16 chr	14,500		Amselem *et al.* (2011)
Botrytis cinerea	a	necrotroph	38	Sanger	118 sc	16,400		Amselem *et al.* (2011)
Fusarium oxysporum	a	saprotroph	60	WGS	114 sc	17,735	lineage-specific chr.	Ma *et al.* (2010)
Neurospora crassa	a	saprotroph	40	WGS	958 sc	10,000	model organism	Galagan *et al.* (2003)

[a] a=ascomycete, b=basidiomycete,o=oomycete
[b] WGS=whole genome shotgun sequencing, p=plasmids, f=fosmids, PE=paired end reads
[c] ch=chromosomes, sc=scaffolds, ctg=contigs

References

N. Agmon, S. Pur, B. Liefshitz and M. Kupiec (2009) Analysis of repair mechanism choice during homologous recombination. *Nucleic Acids Res*, **37**, 5081–5092.

E.D. Akhunov, A.W. Goodyear, S. Geng, L.L. Qi, B. Echalier, B.S. Gill, Miftahudin, J.P. Gustafson, G. Lazo, S. Chao, O.D. Anderson, A.M. Linkiewicz, J. Dubcovsky, M.L. Rota, M.E. Sorrells, D. Zhang, H.T. Nguyen, V. Kalavacharla, K. Hossain, S.F. Kianian, J. Peng, N.L.V. Lapitan, J.L. Gonzalez-Hernandez, J.A. Anderson, D.W. Choi, T.J. Close, M. Dilbirligi, K.S. Gill, M.K. Walker-Simmons, C. Steber, P.E. McGuire, C.O. Qualset and J. Dvorak (2003) The organization and rate of evolution of wheat genomes are correlated with recombination rates along chromosome arms. *Genome Res*, **13**, 753–763.

S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

J. Amselem, C.A. Cuomo, J.A.L. van Kan, M. Viaud, E.P. Benito, A. Couloux, P.M. Coutinho, R.P. de Vries, P.S. Dyer, S. Fillinger, E. Fournier, L. Gout, M. Hahn, L. Kohn, N. Lapalu, K.M. Plummer, J.M. Pradier, E. Quévillon, A. Sharon, A. Simon, A. ten Have, B. Tudzynski, P. Tudzynski, P. Wincker, M. Andrew, V. Anthouard, R.E. Beever, R. Beffa, I. Benoit, O. Bouzid, B. Brault, Z. Chen, M. Choquer, J. Collémare, P. Cotton, E.G. Danchin, C.D. Silva, A. Gautier, C. Giraud, T. Giraud, C. Gonzalez, S. Grossetete, U. Güldener, B. Henrissat, B.J. Howlett, C. Kodira, M. Kretschmer, A. Lappartient, M. Leroch, C. Levis, E. Mauceli, C. Neuvéglise, B. Oeser, M. Pearson, J. Poulain, N. Poussereau, H. Quesneville, C. Rascle, J. Schumacher, B. Ségurens, A. Sexton, E. Silva, C. Sirven, D.M. Soanes, N.J. Talbot, M. Templeton, C. Yandava, O. Yarden, Q. Zeng, J.A. Rollins, M.H. Lebrun and M. Dickman (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genet*, **7**, e1002230.

L. Baxter, S. Tripathy, N. Ishaque, N. Boot, A. Cabral, E. Kemen, M. Thines, A. Ah-Fong, R. Anderson, W. Badejoko, P. Bittner-Eddy, J.L. Boore, M.C. Chibucos, M. Coates, P. Dehal, K. Delehaunty, S. Dong, P. Downton, B. Dumas, G. Fabro, C. Fronick, S.I. Fuerstenberg, L. Fulton, E. Gaulin, F. Govers, L. Hughes, S. Humphray, R.H.Y. Jiang, H. Judelson, S. Kamoun, K. Kyung, H. Meijer, P. Minx, P. Morris, J. Nelson, V. Phuntumart, D. Qutob, A. Rehmany, A. Rougon-Cardoso, P. Ryden, T. Torto-Alalibo, D. Studholme, Y. Wang, J. Win, J. Wood, S.W. Clifton, J. Rogers, G.V. den Ackerveken, J.D.G. Jones, J.M. McDowell, J. Beynon and B.M. Tyler (2010) Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science*, **330**, 1549–1551.

J.L. Bennetzen, C. Coleman, R. Liu, J. Ma and W. Ramakrishna (2004) Consistent over-estimation of gene number in complex plant genomes. *Curr Opin Plant Biol*, **7**, 732–736.

N.K. Bhullar, Z. Zhang, T. Wicker and B. Keller (2010) Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene *Pm3*: a large scale allele mining project. *BMC Plant Biol*, **10**, 88.

L.V. Bindschedler, T.A. Burgis, D.J.S. Mills, J.T.C. Ho, R. Cramer and P.D. Spanu (2009) In planta proteomics and proteogenomics of the biotrophic barley fungal pathogen *Blumeria graminis* f. sp. *hordei*. *Mol Cell Proteomics*, **8**, 2368–2381.

L. Borbye, I. Linde-Laursen, S.K. Christiansen and H. Giese (1992) The chromosome complement of *Erysiphe graminis* f.sp. *hordei* analysed by light microscopy and fiel inversion gel electrophoresis. *Mycol Res*, **96**, 97–102.

M.E. Bougnoux, C. Pujol, D. Diogo, C. Bouchier, D.R. Soll and C. d'Enfert (2008) Mating is rare within as well as between clades of the human pathogen *Candida albicans*. *Fungal Genet Biol*, **45**, 221–231.

U. Braun (2011) The current systematics and taxonomy of the powdery mildews (Erysiphales): an overview. *Mycoscience*, **52**, 210–212.

M.T. Brewer, L. Cadle-Davidson, P. Cortesi, P.D. Spanu and M.G. Milgroom (2011) Identification and structure of the mating-type locus and development of PCR-based markers for mating type in powdery mildew fungi. *Fungal Genet Biol*, **48**, 704–713.

S. Brunner, S. Hurni, P. Streckeisen, G. Mayr, M. Albrecht, N. Yahiaoui and B. Keller (2010) Intragenic allele pyramiding combines different specificities of wheat *Pm3* resistance alleles. *Plant J*, **64**, 433–445.

J.P. Buchmann, T. Matsumoto, N. Stein, B. Keller and T. Wicker (2012) Inter-species sequence comparison of *Brachypodium* reveals how transposon activity corrodes genome colinearity. *Plant J*, **71**, 550–563.

R. Büschges, K. Hollricher, R. Panstruga, G. Simons, M. Wolter, A. Frijters, R. van Daelen, T. van der Lee, P. Diergaarde, J. Groenendijk, S. Töpsch, P. Vos, F. Salamini and P. Schulze-Lefert (1997) The barley *Mlo* gene: a novel control element of plant pathogen resistance. *Cell*, **88**, 695–705.

D. Cantu, M. Govindarajulu, A. Kozik, M. Wang, X. Chen, K.K. Kojima, J. Jurka, R.W. Michelmore and J. Dubcovsky (2011) Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PLoS One*, **6**, e24230.

A. Cao, L. Xing, X. Wang, X. Yang, W. Wang, Y. Sun, C. Qian, J. Ni, Y. Chen, D. Liu, X. Wang and P. Chen (2011) Serine/threonine kinase gene Stpk-V, a key member of powdery mildew resistance gene *Pm21*, confers powdery mildew resistance in wheat. *Proc Natl Acad Sci U S A*, **108**, 7727–7732.

D. Chalupska, H.Y. Lee, J.D. Faris, A. Evrard, B. Chalhoub, R. Haselkorn and P. Gornicki (2008) Acc homoeoloci and the evolution of wheat genomes. *Proc Natl Acad Sci U S A*, **105**, 9691–9696.

Y.L. Chang, S. Cho, H.C. Kistler, C.S. Hsieh and G.J. Muehlbauer (2007) Bacterial artificial chromosome-based physical map of *Gibberella zeae* (*Fusarium graminearum*). *Genome*, **50**, 954–962.

J.M. Chen, D.N. Cooper, N. Chuzhanova, C. FÃl'rec and G.P. Patrinos (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*, **8**, 762–775.

A. Conesa, S. Götz, J.M. García-Gómez, J. Terol, M. Talón and M. Robles (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674–3676.

R. Conner, A. Kuzyk and H. Su (2003) Impact of powdery mildew on the yield of soft white spring wheat cultivars. *Canadian Journal of Plant Science*, **83**, 725–728.

I.H.G.S. Consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

E. Coppin, R. Debuchy, S. Arnaise and M. Picard (1997) Mating types and sexual development in filamentous ascomycetes. *Microbiol Mol Biol Rev*, **61**, 411–428.

B.C. Couch and L.M. Kohn (2002) A multilocus gene genealogy concordant with host preference indicates segregation of a new species, Magnaporthe oryzae, from M. grisea. *Mycologia*, **94**, 683–693.

C.A. Cuomo, U. Güldener, J.R. Xu, F. Trail, B.G. Turgeon, A.D. Pietro, J.D. Walton, L.J. Ma, S.E. Baker, M. Rep, G. Adam, J. Antoniw, T. Baldwin, S. Calvo, Y.L. Chang, D. Decaprio, L.R. Gale, S. Gnerre, R.S. Goswami, K. Hammond-Kosack, L.J. Harris, K. Hilburn, J.C. Kennell, S. Kroken, J.K. Magnuson, G. Mannhaupt, E. Mauceli, H.W. Mewes, R. Mitterbauer, G. Muehlbauer, M. Münsterkötter, D. Nelson, K. O'donnell, T. Ouellet, W. Qi, H. Quesneville, M.I.G. Roncero, K.Y. Seong, I.V. Tetko, M. Urban, C. Waalwijk, T.J. Ward, J. Yao, B.W. Birren and H.C. Kistler (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.

B. Curtis, S. Rajaram and H.G. Macpherson (2002) Bread wheat improvement and production. *Plant Production and Protection*, **Series 567 ISBN: 9251048096**, Y4011/E.

R. Dean, J.A.L.V. Kan, Z.A. Pretorius, K.E. Hammond-Kosack, A.D. Pietro, P.D. Spanu, J.J. Rudd, M. Dickman, R. Kahmann, J. Ellis and G.D. Foster (2012) The Top 10 fungal pathogens in molecular plant pathology. *Mol Plant Pathol*, **13**, 414–430.

R.A. Dean, N.J. Talbot, D.J. Ebbole, M.L. Farman, T.K. Mitchell, M.J. Orbach, M. Thon, R. Kulkarni, J.R. Xu, H. Pan, N.D. Read, Y.H. Lee, I. Carbone, D. Brown, Y.Y. Oh, N. Donofrio, J.S. Jeong, D.M. Soanes, S. Djonovic, E. Kolomiets, C. Rehmeyer, W. Li, M. Harding, S. Kim, M.H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L.J. Ma, R. Nicol, S. Purcell, C. Nusbaum, J.E. Galagan and B.W. Birren (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, **434**, 980–986.

R. Debuchy and B. Turgeon (2006) *The Mycota: Growth, Differentiation, and Sexuality*, chapter Mating-type structure, function and evolution in Euascomycetes, pages 293–323. Springer-Verlag, Berlin.

P.J.G.M. DeWit, R. Mehrabi, H.A.V. den Burg and I. Stergiopoulos (2009) Fungal effector proteins: past, present and future. *Mol Plant Pathol*, **10**, 735–747.

A. Djamei, K. Schipper, F. Rabe, A. Ghosh, V. Vincon, J. Kahnt, S. Osorio, T. Tohge, A.R. Fernie, I. Feussner, K. Feussner, P. Meinicke, Y.D. Stierhof, H. Schwarz, B. Macek, M. Mann and R. Kahmann (2011) Metabolic priming by a secreted fungal effector. *Nature*, **478**, 395–398.

P.N. Dodds, G.J. Lawrence, A.M. Catanzariti, M.A. Ayliffe and J.G. Ellis (2004) The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *Plant Cell*, **16**, 755–768.

P.N. Dodds, G.J. Lawrence, A.M. Catanzariti, T. Teh, C.I.A. Wang, M.A. Ayliffe, B. Kobe and J.G. Ellis (2006) Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proc Natl Acad Sci U S A*, **103**, 8888–8893.

G. Doehlemann, K. van der Linde, D. Assmann, D. Schwammbach, A. Hof, A. Mohanty, D. Jackson and R. Kahmann (2009) *Pep1*, a secreted effector protein of *Ustilago maydis*, is required for successful invasion of plant cells. *PLoS Pathog*, **5**, e1000290.

S. Dong, W. Yin, G. Kong, X. Yang, D. Qutob, Q. Chen, S.D. Kale, Y. Sui, Z. Zhang, D. Dou, X. Zheng, M. Gijzen, B.M. Tyler and Y. Wang (2011) *Phytophthora sojae* avirulence effector *Avr3b* is a secreted NADH and ADP-ribose pyrophosphorylase that modulates plant immunity. *PLoS Pathog*, **7**, e1002353.

D. Dou, S.D. Kale, X. Wang, Y. Chen, Q. Wang, X. Wang, R.H.Y. Jiang, F.D. Arredondo, R.G. Anderson, P.B. Thakur, J.M. McDowell, Y. Wang and B.M. Tyler (2008) Conserved C-terminal motifs required for avirulence and suppression of cell death by *Phytophthora sojae* effector *Avr1b*. *Plant Cell*, **20**, 1118–1133.

J. Draper, L.A. Mur, G. Jenkins, G.C. Ghosh-Biswas, P. Bablak, R. Hasterok and A.P. Routledge (2001) *Brachypodium distachyon*. A new model system for functional genomics in grasses. *Plant Physiol*, **127**, 1539–1555.

S. Duplessis, C.A. Cuomo, Y.C. Lin, A. Aerts, E. Tisserant, C. Veneault-Fourrey, D.L. Joly, S. Hacquard, J. Amselem, B.L. Cantarel, R. Chiu, P.M. Coutinho, N. Feau, M. Field, P. Frey, E. Gelhaye, J. Goldberg, M.G. Grabherr, C.D. Kodira, A. Kohler, U. Kües, E.A. Lindquist, S.M. Lucas, R. Mago, E. Mauceli, E. Morin, C. Murat, J.L. Pangilinan, R. Park, M. Pearson, H. Quesneville, N. Rouhier, S. Sakthikumar, A.A. Salamov, J. Schmutz, B. Selles, H. Shapiro, P. Tanguay, G.A. Tuskan, B. Henrissat, Y.V. de Peer, P. Rouzé, J.G. Ellis, P.N. Dodds, J.E. Schein, S. Zhong, R.C. Hamelin, I.V. Grigoriev, L.J. Szabo and F. Martin (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc Natl Acad Sci U S A*, **108**, 9166–9171.

J. Dvorak and E.D. Akhunov (2005) Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the *Aegilops-Triticum* alliance. *Genetics*, **171**, 323–332.

N. Eshed, A. Dinoor and Y. Litwin (1994) The physiological specialization of wheat powdery mildew in Israel and the search for mildew resistance in wild wheat *Triticum dicoccoides*. *Phytoparasitica*, **22**, 49–90.

N. Eshed and I. Wahl (1970) Host ranges and interrelations of *Erysiphe graminis hordei*, *Erysiphe graminis tritici*, and *Erysiphe graminis avenae*. *Phytopathology*, **60**, 628–634.

N. Eshed and I. Wahl (1975) Role of wild grasses in epidemics of powdery mildew in small grains in Israel. *Phytopathology*, **65**, 57–63.

B. Ewing, L. Hillier, M.C. Wendl and P. Green (1998) Base-calling of automated sequencer traces using phred. I. accuracy assessment. *Genome Res*, **8**, 175–185.

H.H. Flor (1971) Current status of the gene-for-gene concept. *Ann Rev Phytopathol*, **9**, 275–296.

T. Flutre, E. Duprat, C. Feuillet and H. Quesneville (2011) Considering transposable element diversification in *de novo* annotation approaches. *PLoS One*, **6**, e16526.

Z. Frenkel, E. Paux, D. Mester, C. Feuillet and A. Korol (2010) LTC: a novel algorithm to improve the efficiency of contig assembly for physical mapping in complex genomes. *BMC Bioinformatics*, **11**, 584.

J.E. Galagan, S.E. Calvo, K.A. Borkovich, E.U. Selker, N.D. Read, D. Jaffe, W. FitzHugh, L.J. Ma, S. Smirnov, S. Purcell, B. Rehman, T. Elkins, R. Engels, S. Wang, C.B. Nielsen, J. Butler, M. Endrizzi, D. Qui, P. Ianakiev, D. Bell-Pedersen, M.A. Nelson, M. Werner-Washburne, C.P. Selitrennikoff, J.A. Kinsey, E.L. Braun, A. Zelter, U. Schulte, G.O. Kothe, G. Jedd, W. Mewes, C. Staben, E. Marcotte, D. Greenberg, A. Roy, K. Foley, J. Naylor, N. Stange-Thomann, R. Barrett, S. Gnerre, M. Kamal, M. Kamvysselis, E. Mauceli, C. Bielke, S. Rudd, D. Frishman, S. Krystofova, C. Rasmussen, R.L. Metzenberg, D.D. Perkins, S. Kroken, C. Cogoni, G. Macino, D. Catcheside, W. Li, R.J. Pratt, S.A. Osmani, C.P.C. DeSouza, L. Glass, M.J. Orbach, J.A. Berglund, R. Voelker, O. Yarden, M. Plamann, S. Seiler, J. Dunlap, A. Radford, R. Aramayo, D.O. Natvig, L.A. Alex, G. Mannhaupt, D.J. Ebbole, M. Freitag, I. Paulsen, M.S. Sachs, E.S. Lander, C. Nusbaum and B. Birren (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*, **422**, 859–868.

B.S. Gaut, B.R. Morton, B.C. McCaig and M.T. Clegg (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc Natl Acad Sci U S A*, **93**, 10274–10279.

D.A. Glawe (2008) The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. *Annu Rev Phytopathol*, **46**, 27–51.

D. Godfrey, H. Böhlenius, C. Pedersen, Z. Zhang, J. Emmersen and H. Thordal-Christensen (2010) Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. *BMC Genomics*, **11**, 317.

D. Godfrey, Z. Zhang, G. Saalbach and H. Thordal-Christensen (2009) A proteomics study of barley powdery mildew haustoria. *Proteomics*, **9**, 3222–3232.

S.B. Goodwin, S.B. M'barek, B. Dhillon, A.H.J. Wittenberg, C.F. Crane, J.K. Hane, A.J. Foster, T.A.J.V. der Lee, J. Grimwood, A. Aerts, J. Antoniw, A. Bailey, B. Bluhm, J. Bowler, J. Bristow, A. van der Burgt, B. Canto-Canché, A.C.L. Churchill, L. Conde-Ferràez, H.J. Cools, P.M. Coutinho, M. Csukai, P. Dehal, P.D. Wit, B. Donzelli, H.C. van de Geest, R.C.H.J. van Ham, K.E. Hammond-Kosack, B. Henrissat, A. Kilian, A.K. Kobayashi, E. Koopmann, Y. Kourmpetis, A. Kuzniar, E. Lindquist, V. Lombard, C. Maliepaard, N. Martins, R. Mehrabi, J.P.H. Nap, A. Ponomarenko, J.J. Rudd, A. Salamov, J. Schmutz, H.J. Schouten, H. Shapiro, I. Stergiopoulos, S.F.F. Torriani, H. Tu, R.P. de Vries, C. Waalwijk, S.B. Ware, A. Wiebenga, L.H. Zwiers, R.P. Oliver, I.V. Grigoriev and G.H.J. Kema (2011) Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet*, **7**, e1002070.

T.R. Gregory, J.A. Nicol, H. Tamm, B. Kullman, K. Kullman, I.J. Leitch, B.G. Murray, D.F. Kapraun, J. Greilhuber and M.D. Bennett (2007) Eukaryotic genome size databases. *Nucleic Acids Res*, **35**, D332–D338.

Griffey (1993) Effectiveness of adult-plant resistance in reducing grain yield loss to powdery mildew in winter wheat. *Plant Disease*, **77**, 618–622.

B.J. Haas, S. Kamoun, M.C. Zody, R.H.Y. Jiang, R.E. Handsaker, L.M. Cano, M. Grabherr, C.D. Kodira, S. Raffaele, T. Torto-Alalibo, T.O. Bozkurt, A.M.V. Ah-Fong, L. Alvarado, V.L. Anderson, M.R. Armstrong, A. Avrova, L. Baxter, J. Beynon, P.C. Boevink, S.R. Bollmann, J.I.B. Bos, V. Bulone, G. Cai, C. Cakir, J.C. Carrington, M. Chawner, L. Conti, S. Costanzo, R. Ewan, N. Fahlgren, M.A. Fischbach, J. Fugelstad, E.M. Gilroy, S. Gnerre, P.J. Green, L.J. Grenville-Briggs, J. Griffith, N.J. Grünwald, K. Horn, N.R. Horner, C.H. Hu, E. Huitema, D.H. Jeong, A.M.E. Jones, J.D.G. Jones, R.W. Jones, E.K. Karlsson, S.G. Kunjeti, K. Lamour, Z. Liu, L. Ma, D. Maclean, M.C. Chibucos, H. McDonald, J. McWalters, H.J.G. Meijer, W. Morgan, P.F. Morris, C.A. Munro, K. O'Neill, M. Ospina-Giraldo, A. Pinzón, L. Pritchard, B. Ramsahoye, Q. Ren, S. Restrepo, S. Roy, A. Sadanandom, A. Savidor, S. Schornack, D.C. Schwartz, U.D. Schumann, B. Schwessinger, L. Seyer, T. Sharpe, C. Silvar, J. Song, D.J. Studholme, S. Sykes, M. Thines, P.J.I. van de Vondervoort, V. Phuntumart, S. Wawra, R. Weide, J. Win, C. Young, S. Zhou, W. Fry, B.C. Meyers, P. van West, J. Ristaino, F. Govers, P.R.J. Birch, S.C. Whisson, H.S. Judelson and C. Nusbaum (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*, **461**, 393–398.

D. Halterman, F. Zhou, F. Wei, R.P. Wise and P. Schulze-Lefert (2001) The MLA6 coiled-coil, NBS-LRR protein confers *AvrMla6*-dependent resistance specificity to *Blumeria graminis* f. sp. *hordei* in barley and wheat. *Plant J*, **25**, 335–348.

J.K. Hane, R.G.T. Lowe, P.S. Solomon, K.C. Tan, C.L. Schoch, J.W. Spatafora, P.W. Crous, C. Kodira, B.W. Birren, J.E. Galagan, S.F.F. Torriani, B.A. McDonald and R.P. Oliver (2007) Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell*, **19**, 3347–3368.

J. Hardison (1944) *Phytopathology*, **34**, 1–20.

A. Haudry, A. Cenci, C. Ravel, T. Bataillon, D. Brunel, C. Poncet, I. Hochu, S. Poirier, S. Santoni, S. Glémin and J. David (2007) Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol Biol Evol*, **24**, 1506–1517.

D. Hawksworth (1991) The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycological Research*, **95**, 641–655.

J. Heitman (2006) Sexual reproduction and the evolution of microbial pathogens. *Curr Biol*, **16**, R711–R725.

Hiura (1978) *The Powdery Mildews*, chapter Genetic basis of *formae speciales*., pages 101–128. Academic Press, New York.

Hsam and Zeller (2002) *The powdery mildews: a comprehensive treatise*, chapter 14, pages 219–238. APS press.

S. Huang, A. Sirikhachornkit, J.D. Faris, X. Su, B.S. Gill, R. Haselkorn and P. Gornicki (2002) Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol Biol*, **48**, 805–820.

R. Hückelhoven and R. Panstruga (2011) Cell biology of the plant-powdery mildew interaction. *Curr Opin Plant Biol*, **14**, 738–746.

L.A. Inda, J.G. Segarra-Moragues, J. Müller, P.M. Peterson and P. Catalán (2008) Dated historical biogeography of the temperate Loliinae (Poaceae, Pooideae) grasses in the northern and southern hemispheres. *Mol Phylogenet Evol*, **46**, 932–957.

I.B. Initiative (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

T. Inuma, S.A. Khodaparast and S. Takamatsu (2007) Multilocus phylogenetic analyses within *Blumeria graminis*, a powdery mildew fungus of cereals. *Mol Phylogenet Evol*, **44**, 741–751.

J.P. Javerzat, V. Bhattacherjee and C. Barreau (1993) Isolation of telomeric DNA from the filamentous fungus *Podospora anserina* and construction of a self-replicating linear plasmid showing high transformation frequency. *Nucleic Acids Res*, **21**, 497–504.

J.D.G. Jones and J.L. Dangl (2006) The plant immune system. *Nature*, **444**, 323–329.

I.H. Jørgensen (1992) Discovery, characterization and exploitation of *Mlo* powdery mildew resistance in barley. *Euphytica*, **63**, 141–152.

P.F. K. Everts S. Leath (2001) Impact of powdery mildew and leaf rust on milling and baking quality of soft red winter wheat. *Plant Dis.*, **85**, 423–429.

J. Kämper, R. Kahmann, M. Bölker, L.J. Ma, T. Brefort, B.J. Saville, F. Banuett, J.W. Kronstad, S.E. Gold, O. Müller, M.H. Perlin, H.A.B. Wösten, R. de Vries, J. Ruiz-Herrera, C.G. Reynaga-Peña, K. Snetselaar, M. McCann, J. Perez-Martín, M. Feldbrügge, C.W. Basse, G. Steinberg, J.I. Ibeas, W. Holloman, P. Guzman, M. Farman, J.E. Stajich, R. Sentandreu, J.M. González-Prieto, J.C. Kennell, L. Molina, J. Schirawski, A. Mendoza-Mendoza, D. Greilinger, K. Münch, N. Rössel,

M. Scherer, M. Vranes, O. Ladendorf, V. Vincon, U. Fuchs, B. Sandrock, S. Meng, E.C.H. Ho, M.J. Cahill, K.J. Boyce, J. Klose, S.J. Klosterman, H.J. Deelstra, L. Ortiz-Castellanos, W. Li, P. Sanchez-Alonso, P.H. Schreier, I. Häuser-Hahn, M. Vaupel, E. Koopmann, G. Friedrich, H. Voss, T. Schlüter, J. Margolis, D. Platt, C. Swimmer, A. Gnirke, F. Chen, V. Vysotskaia, G. Mannhaupt, U. Güldener, M. Münsterkötter, D. Haase, M. Oesterheld, H.W. Mewes, E.W. Mauceli, D. DeCaprio, C.M. Wade, J. Butler, S. Young, D.B. Jaffe, S. Calvo, C. Nusbaum, J. Galagan and B.W. Birren (2006) Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature*, **444**, 97–101.

T. Kasuga, T.J. White and J.W. Taylor (2002) Estimation of nucleotide substitution rates in Eurotiomycete fungi. *Mol Biol Evol*, **19**, 2318–2324.

K. Kazan, D.M. Gardiner and J.M. Manners (2012) On the trail of a cereal killer: recent advances in *Fusarium graminearum* pathogenomics and host resistance. *Mol Plant Pathol*, **13**, 399–413.

E. Kemen, A. Gardiner, T. Schultz-Larsen, A.C. Kemen, A.L. Balmuth, A. Robert-Seilaniantz, K. Bailey, E. Holub, D.J. Studholme, D. Maclean and J.D.G. Jones (2011) Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol*, **9**, e1001094.

E. Kemen and J.D.G. Jones (2012) Obligate biotroph parasitism: can we link genomes to lifestyles? *Trends Plant Sci*, **17**, 448–457.

D.U. Kim, J. Hayles, D. Kim, V. Wood, H.O. Park, M. Won, H.S. Yoo, T. Duhig, M. Nam, G. Palmer, S. Han, L. Jeffery, S.T. Baek, H. Lee, Y.S. Shim, M. Lee, L. Kim, K.S. Heo, E.J. Noh, A.R. Lee, Y.J. Jang, K.S. Chung, S.J. Choi, J.Y. Park, Y. Park, H.M. Kim, S.K. Park, H.J. Park, E.J. Kang, H.B. Kim, H.S. Kang, H.M. Park, K. Kim, K. Song, K.B. Song, P. Nurse and K.L. Hoe (2010) Analysis of a genome-wide set of gene deletions in the fission yeast *Schizosaccharomyces pombe*. *Nat Biotechnol*, **28**, 617–623.

M. Kimura (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*, **16**, 111–120.

S.A. Lee, S. Wormsley, S. Kamoun, A.F.S. Lee, K. Joiner and B. Wong (2003) An analysis of the *Candida albicans* genome database for soluble secreted proteins using computer-based prediction algorithms. *Yeast*, **20**, 595–610.

M.G. Links, E. Holub, R.H.Y. Jiang, A.G. Sharpe, D. Hegedus, E. Beynon, D. Sillito, W.E. Clarke, S. Uzuhashi and M.H. Borhan (2011) *De novo* sequence assembly of *Albugo candida* reveals a small genome relative to other biotrophic oomycetes. *BMC Genomics*, **12**, 503.

N. Liu, G. Gong, M. Zhang, Y. Zhou, Z. Chen, J. Yang, H. Chen, X. Wang, Y. Lei and K. Liu (2012) Over-summering of wheat powdery mildew in Sichuan Province, China. *Crop Protection*, **34**, 112–118.

J. Ma and J.L. Bennetzen (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A*, **101**, 12404–12410.

L.J. Ma, H.C. van der Does, K.A. Borkovich, J.J. Coleman, M.J. Daboussi, A.D. Pietro, M. Dufresne, M. Freitag, M. Grabherr, B. Henrissat, P.M. Houterman, S. Kang, W.B. Shim, C. Woloshuk, X. Xie, J.R. Xu, J. Antoniw, S.E. Baker, B.H. Bluhm, A. Breakspear, D.W. Brown, R.A.E. Butchko, S. Chapman, R. Coulson, P.M. Coutinho, E.G.J. Danchin, A. Diener, L.R. Gale, D.M. Gardiner, S. Goff, K.E. Hammond-Kosack, K. Hilburn, A. Hua-Van, W. Jonkers, K. Kazan, C.D. Kodira, M. Koehrsen, L. Kumar, Y.H. Lee, L. Li, J.M. Manners, D. Miranda-Saavedra, M. Mukherjee, G. Park, J. Park, S.Y. Park, R.H. Proctor, A. Regev, M.C. Ruiz-Roldan, D. Sain, S. Sakthikumar, S. Sykes, D.C. Schwartz, B.G. Turgeon, I. Wapinski, O. Yoder, S. Young, Q. Zeng, S. Zhou, J. Galagan, C.A. Cuomo, H.C. Kistler and M. Rep (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*, **464**, 367–373.

E. Marchal (1902) De la specialisation du parisitisme chez l'*Erysiphe graminis*. *Compt Rend Acad Sci Paris*, **135**, 210–212.

F. Martin, A. Aerts, D. Ahrén, A. Brun, E.G.J. Danchin, F. Duchaussoy, J. Gibon, A. Kohler, E. Lindquist, V. Pereda, A. Salamov, H.J. Shapiro, J. Wuyts, D. Blaudez, M. Buée, P. Brokstein, B. Canbäck, D. Cohen, P.E. Courty, P.M. Coutinho, C. Delaruelle, J.C. Detter, A. Deveau, S. DiFazio, S. Duplessis, L. Fraissinet-Tachet, E. Lucic, P. Frey-Klett, C. Fourrey, I. Feussner, G. Gay, J. Grimwood, P.J. Hoegger, P. Jain, S. Kilaru, J. Labbé, Y.C. Lin, V. Legué, F.L. Tacon, R. Marmeisse, D. Melayah, B. Montanini, M. Muratet, U. Nehls, H. Niculita-Hirzel, M.P.O.L. Secq, M. Peter, H. Quesneville, B. Rajashekar, M. Reich, N. Rouhier, J. Schmutz, T. Yin, M. Chalot, B. Henrissat, U. Kües, S. Lucas, Y.V. de Peer, G.K. Podila, A. Polle, P.J. Pukkila, P.M. Richardson, P. Rouzé, I.R. Sanders, J.E. Stajich, A. Tunlid, G. Tuskan and I.V. Grigoriev (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature*, **452**, 88–92.

F. Martin, A. Kohler, C. Murat, R. Balestrini, P.M. Coutinho, O. Jaillon, B. Montanini, E. Morin, B. Noel, R. Percudani, B. Porcel, A. Rubini, A. Amicucci, J. Amselem, V. Anthouard, S. Arcioni, F. Artiguenave, J.M. Aury, P. Ballario, A. Bolchi, A. Brenna, A. Brun, M. Buée, B. Cantarel, G. Chevalier, A. Couloux, C.D. Silva, F. Denoeud, S. Duplessis, S. Ghignone, B. Hilselberger, M. Iotti, B. Marçais, A. Mello, M. Miranda, G. Pacioni, H. Quesneville, C. Riccioni, R. Ruotolo, R. Splivallo, V. Stocchi, E. Tisserant, A.R. Viscomi, A. Zambonelli, E. Zampieri, B. Henrissat, M.H. Lebrun, F. Paolocci, P. Bonfante, S. Ottonello and P. Wincker (2010a) Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature*, **464**, 1033–1038.

T. Martin, S.W. Lu, H. van Tilbeurgh, D.R. Ripoll, C. Dixelius, B.G. Turgeon and R. Debuchy (2010b) Tracing the origin of the fungal a1 domain places its ancestor in the HMG-box superfamily: implication for fungal mating-type evolution. *PLoS One*, **5**, e15199.

J.H. McDonald and M. Kreitman (1991) Adaptive protein evolution at the *Adh* locus in Drosophila. *Nature*, **351**, 652–654.

M.L. Metzker (2010) Sequencing technologies - the next generation. *Nat Rev Genet*, **11**, 31–46.

R.W. Michelmore and B.C. Meyers (1998) Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. *Genome Res*, **8**, 1113–1130.

E. Mühle and K. Frauenstein (1962a) Untersuchungen zur physiologischen Spezialisierung von *Erysiphe graminis* DC. *Der Züchter*, **32**, 324–327.

E. Mühle and K. Frauenstein (1962b) Untersuchungen zur physiologischen Spezialisierung von *Erysiphe graminis* DC. *Der Züchter*, **32**, 345–352.

E. Mühle and K. Frauenstein (1963) Untersuchungen zur physiologischen Spezialisierung von *Erysiphe graminis* DC. *Der Züchter*, **33**, 124–131.

E. Mühle and K. Frauenstein (1970) Untersuchungen zur physiologischen Spezialisierung von *Erysiphe graminis* DC. *Theor Appl Genet*, **40**, 56–58.

S. Noir, T. Colby, A. Harzen, J. Schmidt and R. Panstruga (2009) A proteomic analysis of powdery mildew (*Blumeria graminis* f.sp. *hordei*) conidiospores. *Mol Plant Pathol*, **10**, 223–236.

C. Nombela, C. Gil and W.L. Chaffin (2006) Non-conventional protein secretion in yeast. *Trends Microbiol*, **14**, 15–21.

D. Nowara, A. Gay, C. Lacomme, J. Shaw, C. Ridout, D. Douchkov, G. Hensel, J. Kumlehn and P. Schweizer (2010) HIGS: host-induced gene silencing in the obligate biotrophic fungal pathogen *Blumeria graminis*. *Plant Cell*, **22**, 3130–3141.

M. Nowrousian, J.E. Stajich, M. Chu, I. Engh, E. Espagne, K. Halliday, J. Kamerewerd, F. Kempken, B. Knab, H.C. Kuo, H.D. Osiewacz, S. Pöggeler, N.D. Read, S. Seiler, K.M. Smith, D. Zickler, U. Kück and M. Freitag (2010) *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. *PLoS Genet*, **6**, e1000891.

S. Oberhaensli, F. Parlange, J.P. Buchmann, F.H. Jenny, J.C. Abbott, T.A. Burgis, P.D. Spanu, B. Keller and T. Wicker (2011) Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene colinearity, dates divergence and indicates host-pathogen co-evolution. *Fungal Genet Biol*, **48**, 327–334.

R.J. O'Connell, M.R. Thon, S. Hacquard, S.G. Amyotte, J. Kleemann, M.F. Torres, U. Damm, E.A. Buiate, L. Epstein, N. Alkan, J. Altmüller, L. Alvarado-Balderrama, C.A. Bauser, C. Becker, B.W. Birren, Z. Chen, J. Choi, J.A. Crouch, J.P. Duvick, M.A. Farman, P. Gan, D. Heiman, B. Henrissat, R.J. Howard, M. Kabbage, C. Koch, B. Kracher, Y. Kubo, A.D. Law, M.H. Lebrun, Y.H. Lee, I. Miyara, N. Moore, U. Neumann, K. NordstrÃűm, D.G. Panaccione, R. Panstruga, M. Place, R.H. Proctor, D. Prusky, G. Rech, R. Reinhardt, J.A. Rollins, S. Rounsley, C.L. Schardl, D.C. Schwartz, N. Shenoy, K. Shirasu, U.R. Sikhakolli, K. Stüber, S.A. Sukno, J.A. Sweigard, Y. Takano, H. Takahara, F. Trail, H.C. van der Does, L.M. Voll, I. Will, S. Young, Q. Zeng, J. Zhang, S. Zhou, M.B. Dickman, P. Schulze-Lefert, E.V.L. van Themaat, L.J. Ma and L.J. Vaillancourt (2012) Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses. *Nat Genet*, **44**, 1060–1065.

T. Oku and T. Tsuchizaki (1993) Compatibility of *Erysiphe graminis* hybrids f. sp. *secalis* x f. sp. *tritici* with wheat lines involving resistance genes from rye. *J Phytopath*, **138**, 77–83.

T. Oku, S. Yamashita, Y. Doi and N. Nishihara (1985) Host range and *forma specialis* of cocksfoot powdery mildew fungus (*Erysiphe graminis* DC) found in Japan. *Ann Phytophathol Soc Jpn*, **51**, 613–615.

K. Olesen, T. Carver and M. Lyngkjær (2003) Fungal suppression of resistance against inappropriate *Blumeria graminis* formae speciales in barley, oat and wheat. *Physiol Mol Plant Pathol*, **62**, 37–50.

R.P. Oliver, T.L. Friesen, J.D. Faris and P.S. Solomon (2012) *Stagonospora nodorum*: from pathology to genomics and host resistance. *Annu Rev Phytopathol*, **50**, 23–43.

R. Parks, I. Carbone, J.P. Murphy and C. Cowger (2009) Population genetic analysis of an Eastern U.S. wheat powdery mildew population reveals geographic subdivision and recent common ancestry with U.K. and Israeli populations. *Phytopathology*, **99**, 840–849.

F. Parlange, S. Oberhaensli, J. Breen, M. Platzer, S. Taudien, H. Šimková, T. Wicker, J. Doležel and B. Keller (2011) A major invasion of transposable elements accounts for the large size of the *Blumeria graminis* f.sp. *tritici* genome. *Funct Integr Genomics*, **11**, 671–677.

G. Parra, K. Bradnam and I. Korf (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.

C. Pedersen, S.W. Rasmussen and H. Giese (2002a) A genetic map of *Blumeria graminis* based on functional genes, avirulence genes, and molecular markers. *Fungal Genet Biol*, **35**, 235–246.

C. Pedersen, E.V. van Themaat, L.J. McGuffin, J.C. Abbott, T.A. Burgis, G. Barton, L.V. Bindschedler, X. Lu, T. Maekawa, R. Weßling, R. Cramer, H. Thordal-Christensen, R. Panstruga and P.D. Spanu (2012) Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics*, **13**, 694.

C. Pedersen, B. Wu and H. Giese (2002b) A *Blumeria graminis* f.sp. *hordei* BAC library–contig building and microsynteny studies. *Curr Genet*, **42**, 103–113.

J.H. Peng, D. Sun and E. Nevo (2011) Domestication evolution, genetics and genomics in wheat. *Mol Breeding*, **28**, 281–301.

D.A. Petrov (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet*, **17**, 23–28.

R. Pinhasi, J. Fort and A.J. Ammerman (2005) Tracing the origin and spread of agriculture in Europe. *PLoS Biol*, **3**, e410.

H. Quesneville, C.M. Bergman, O. Andrieu, D. Autard, D. Nouaud, M. Ashburner and D. Anxolabehere (2005) Combined evidence annotation of transposable elements in genome sequences. *PLoS Comput Biol*, **1**, 166–175.

S. Raffaele, R.A. Farrer, L.M. Cano, D.J. Studholme, D. MacLean, M. Thines, R.H.Y. Jiang, M.C. Zody, S.G. Kunjeti, N.M. Donofrio, B.C. Meyers, C. Nusbaum and S. Kamoun (2010) Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science*, **330**, 1540–1543.

S. Raffaele and S. Kamoun (2012) Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol*, **10**, 417–430.

M. Rasmussen, L. Rossen and H. Giese (1993) SINE-like properties of a highly repetitive element in the genome of the obligate parasitic fungus *Erysiphe graminis* f.sp. *hordei*. *Mol Gen Genet*, **239**, 298–303.

C.J. Ridout, P. Skamnioti, O. Porritt, S. Sacristan, J.D.G. Jones and J.K.M. Brown (2006) Multiple avirulence paralogues in cereal powdery mildew fungi may contribute to parasite fitness and defeat of plant resistance. *Plant Cell*, **18**, 2402–2414.

C. Ridout and J. Brown (1999) Physical mapping of avirulence genes in the barley powdery mildew pathogen *Erysiphe graminis* f.sp. *hordei* (abstract). In *The First International Powdery Mildew Conference, Palais des Papes, Avignon, France*.

T. Rouxel, J. Grandaubert, J.K. Hane, C. Hoede, A.P. van de Wouw, A. Couloux, V. Dominguez, V. Anthouard, P. Bally, S. Bourras, A.J. Cozijnsen, L.M. Ciuffetti, A. Degrave, A. Dilmaghani, L. Duret, I. Fudal, S.B. Goodwin, L. Gout, N. Glaser, J. Linglin, G.H.J. Kema, N. Lapalu, C.B. Lawrence, K. May, M. Meyer, B. Ollivier, J. Poulain, C.L. Schoch, A. Simon, J.W. Spatafora, A. Stachowiak, B.G. Turgeon, B.M. Tyler, D. Vincent, J. Weissenbach, J. Amselem, H. Quesneville, R.P. Oliver, P. Wincker, M.H. Balesdent and B.J. Howlett (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat Commun*, **2**, 202.

S. Sacristán, M. Vigouroux, C. Pedersen, P. Skamnioti, H. Thordal-Christensen, C. Micali, J.K.M. Brown and C.J. Ridout (2009) Coevolution between a family of parasite virulence effectors and a class of LINE-1 retrotransposons. *PLoS One*, **4**, e7463.

F. Salamini, H. Ozkan, A. Brandolini, R. Schäfer-Pregl and W. Martin (2002) Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet*, **3**, 429–441.

D. Saleh, J. Milazzo, H. Adreit, D. Tharreau and E. Fournier (2012a) Asexual reproduction induces a rapid and permanent loss of sexual reproduction capacity in the rice fungal pathogen Magnaporthe oryzae: results of in vitro experimental evolution assays. *BMC Evol Biol*, **12**, 42.

D. Saleh, P. Xu, Y. Shen, C. Li, H. Adreit, J. Milazzo, V. Ravigné, E. Bazin, J.L. Nottéghem, E. Fournier and D. Tharreau (2012b) Sex at the origin: an Asian population of the rice blast fungus Magnaporthe oryzae reproduces sexually. *Mol Ecol*, **21**, 1330–1344.

F. Sanger, S. Nicklen and A.R. Coulson (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**, 5463–5467.

P. SanMiguel, B.S. Gaut, A. Tikhonov, Y. Nakajima and J.L. Bennetzen (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet*, **20**, 43–45.

P.J. SanMiguel, W. Ramakrishna, J.L. Bennetzen, C.S. Busso and J. Dubcovsky (2002) Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). *Funct Integr Genomics*, **2**, 70–80.

A. Sboner, X.J. Mu, D. Greenbaum, R.K. Auerbach and M.B. Gerstein (2011) The real cost of sequencing: higher than you think! *Genome Biol*, **12**, 125.

S. Scalabrin, M. Morgante and A. Policriti (2009) Automated FingerPrint Background removal: FPB. *BMC Bioinformatics*, **10**, 127.

E.E. Schadt, S. Turner and A. Kasarskis (2010) A window into third-generation sequencing. *Hum Mol Genet*, **19**, R227–R240.

P. Schulze-Lefert and J. Vogel (2000) Closing the ranks to attack by powdery mildew. *Trends Plant Sci*, **5**, 343–348.

S. Seeholzer, T. Tsuchimatsu, T. Jordan, S. Bieri, S. Pajonk, W. Yang, A. Jahoor, K.K. Shimizu, B. Keller and P. Schulze-Lefert (2010) Diversity at the *Mla* powdery mildew resistance locus from cultivated barley reveals sites of positive selection. *Mol Plant Microbe Interact*, **23**, 497–509.

Q.H. Shen, Y. Saijo, S. Mauch, C. Biskup, S. Bieri, B. Keller, H. Seki, B. Ulker, I.E. Somssich and P. Schulze-Lefert (2007) Nuclear activity of MLA immune receptors links isolate-specific and basal disease-resistance responses. *Science*, **315**, 1098–1103.

H. Simková, J. Safár, M. Kubaláková, P. Suchánková, J. Číhalíková, H. Robert-Quatre, P. Azhaguvel, Y. Weng, J. Peng, N.L.V. Lapitan, Y. Ma, F.M. You, M.C. Luo, J. Bartos and J. Dolezel (2011) BAC libraries from wheat chromosome 7D: efficient tool for positional cloning of aphid resistance genes. *J Biomed Biotechnol*, **2011**, 302543.

R.P. Singh, D.P. Hodson, J. Huerta-Espino, Y. Jin, S. Bhavani, P. Njau, S. Herrera-Foessel, P.K. Singh, S. Singh and V. Govindan (2011) The emergence of Ug99 races of the stem rust fungus is a threat to world wheat production. *Annu Rev Phytopathol*, **49**, 465–481.

C. Soderlund, I. Longden and R. Mott (1997) FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci*, **13**, 523–535.

E.L. Sonnhammer and R. Durbin (1995) A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene*, **167**, GC1–G10.

P.D. Spanu (2012) The genomics of obligate (and nonobligate) biotrophs. *Annu Rev Phytopathol*, **50**, 91–109.

P.D. Spanu, J.C. Abbott, J. Amselem, T.A. Burgis, D.M. Soanes, K. Stüber, E.V.L. van Themaat, J.K.M. Brown, S.A. Butcher, S.J. Gurr, M.H. Lebrun, C.J. Ridout, P. Schulze-Lefert, N.J. Talbot, N. Ahmadinejad, C. Ametz, G.R. Barton, M. Benjdia, P. Bidzinski, L.V. Bindschedler, M. Both, M.T. Brewer, L. Cadle-Davidson, M.M. Cadle-Davidson, J. Collemare, R. Cramer, O. Frenkel, D. Godfrey, J. Harriman, C. Hoede, B.C. King, S. Klages, J. Kleemann, D. Knoll, P.S. Koti, J. Kreplak, F.J. López-Ruiz, X. Lu, T. Maekawa, S. Mahanil, C. Micali, M.G. Milgroom, G. Montana, S. Noir, R.J. O'Connell, S. Oberhaensli, F. Parlange, C. Pedersen, H. Quesneville, R. Reinhardt, M. Rott, S. Sacristán, S.M. Schmidt, M. Schön, P. Skamnioti, H. Sommer, A. Stephens, H. Takahara, H. Thordal-Christensen, M. Vigouroux, R. Wessling, T. Wicker and R. Panstruga (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*, **330**, 1543–1546.

P. Srichumpa, S. Brunner, B. Keller and N. Yahiaoui (2005) Allelic series of four powdery mildew resistance genes at the *Pm3* locus in hexaploid bread wheat. *Plant Physiol*, **139**, 885–895.

M. Stanke and S. Waack (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19 Suppl 2**, ii215–ii225.

I. Stergiopoulos and P.J.G.M. de Wit (2009) Fungal effector proteins. *Annu Rev Phytopathol*, **47**, 233–263.

I. Stojiljkovic, J. Bozja and E. Salajsmic (1994) Molecular-cloning of bacterial-DNA in-vivo using a transposable R6k ori and a P1vir phage. *Journal of Bacteriology*, **176**, 1188–1191.

N. Stoletzki and A. Eyre-Walker (2011) Estimation of the neutrality index. *Mol Biol Evol*, **28**, 63–70.

E.H. Stukenbrock, S. Banke, M. Javan-Nikkhah and B.A. McDonald (2007) Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Mol Biol Evol*, **24**, 398–411.

E.H. Stukenbrock, T. Bataillon, J.Y. Dutheil, T.T. Hansen, R. Li, M. Zala, B.A. McDonald, J. Wang and M.H. Schierup (2011) The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Res*, **21**, 2157–2166.

E.H. Stukenbrock, F.G. Jørgensen, M. Zala, T.T. Hansen, B.A. McDonald and M.H. Schierup (2010) Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. *PLoS Genet*, **6**, e1001189.

E.H. Stukenbrock and B.A. McDonald (2008) The origins of plant pathogens in agro-ecosystems. *Annu Rev Phytopathol*, **46**, 75–100.

E.H. Stukenbrock, F.B. Christiansen, T.T. Hansen, J.Y. Dutheil and M.H. Schierup (2012) Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proc Natl Acad Sci U S A*, **109**, 10954–10959.

S. Takamatsu (2004) Estimation of molecular clocks for ITS and 28S rDNA in Erysiphales. *Mycoscience*, **45**, 340–344.

J.W. Taylor and M.L. Berbee (2006) Dating divergences in the Fungal Tree of Life: review and new analyses. *Mycologia*, **98**, 838–849.

M. Tibayrenc and F.J. Ayala (2012) Reproductive clonality of pathogens: A perspective on pathogenic viruses, bacteria, fungi, and parasitic protozoa. *Proc Natl Acad Sci U S A*.

Y. Tosa (1989) Evidence on wheat for gene-for-gene relationship between *formae speciales* of *Erysiphe graminis* and genera of gramineous plants. *Genome*, **32**, 918–924.

Y. Tosa (1992) A model for the evolution of *formae speciales* and races. *Phytopathology*, **82**, 728–730.

Y. Tosa, H. Tamba, K. Tanaka and S. Mayama (2006) Genetic analysis of host species specificity of *magnaporthe oryzae* isolates from rice and wheat. *Phytopathology*, **96**, 480–484.

V. Troch, K. Audenaert, B. Bekaert, M. Höfte and G. Haesaert (2012) Phylogeography and virulence structure of the powdery mildew population on its 'new' host triticale. *BMC Evol Biol*, **12**, 76.

C. Vitte and O. Panaud (2005) LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res*, **110**, 91–107.

A. Walker, A. Bourguennec, J. Confais, G. Morgant and P. Leroux (2010) Evidence of host-range expansion from new powdery mildew (*Blumeria graminis*) infections of triticale (x Triticosecale) in France. *Plant Pathol*, **60**, 207–220.

Y.D. Wei, D.B. Collinge, V. Smedegaard-Petersen and H. Thordal-Christensen (1996) Characterization of the transcript of a new class of retroposon-type repetitive element cloned from the powdery mildew fungus, *Erysiphe graminis*. *Mol Gen Genet*, **250**, 477–482.

S.C. Whisson, P.C. Boevink, L. Moleleki, A.O. Avrova, J.G. Morales, E.M. Gilroy, M.R. Armstrong, S. Grouffaud, P. van West, S. Chapman, I. Hein, I.K. Toth, L. Pritchard and P.R.J. Birch (2007) A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature*, **450**, 115–118.

T. Wicker, J.P. Buchmann and B. Keller (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res*, **20**, 1229–1237.

T. Wicker, F. Sabot, A. Hua-Van, J.L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel and A.H. Schulman (2007) A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*, **8**, 973–982.

T. Wicker, N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z.D. Liu, J. Dubcovsky and B. Keller (2003) Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. *Plant Cell*, **15**, 1186–1197.

A.H.J. Wittenberg, T.A.J. van der Lee, S.B. M'barek, S.B. Ware, S.B. Goodwin, A. Kilian, R.G.F. Visser, G.H.J. Kema and H.J. Schouten (2009) Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. *PLoS One*, **4**, e5863.

T.D. Wu and C.K. Watanabe (2005) GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, **21**, 1859–1875.

R.A. Wyand and J.K.M. Brown (2003) Genetic and forma specialis diversity in *Blumeria graminis* of cereals and its implications for host-pathogen co-evolution. *Mol Plant Pathol*, **4**, 187–198.

G. Xiao, S.H. Ying, P. Zheng, Z.L. Wang, S. Zhang, X.Q. Xie, Y. Shang, R.J.S. Leger, G.P. Zhao, C. Wang and M.G. Feng (2012) Genomic perspectives on the evolution of fungal entomopathogenicity in *Beauveria bassiana*. *Sci Rep*, **2**, 483.

N. Yahiaoui, S. Brunner and B. Keller (2006) Rapid generation of new powdery mildew resistance genes after wheat domestication. *Plant J*, **47**, 85–98.

N. Yahiaoui, P. Srichumpa, R. Dudler and B. Keller (2004) Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene *Pm3b* from hexaploid wheat. *Plant J*, **37**, 528–538.

Z. Yang and R. Nielsen (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol*, **17**, 32–43.

Z. Yang (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–1591.

F.M. You, M.C. Luo, Y.Q. Gu, G.R. Lazo, K. Deal, J. Dvorak and O.D. Anderson (2007) GenoProfiler: batch processing of high-throughput capillary fingerprinting data. *Bioinformatics*, **23**, 240–242.

J. Yu, P.K. Chang, K.C. Ehrlich, J.W. Cary, D. Bhatnagar, T.E. Cleveland, G.A. Payne, J.E. Linz, C.P. Woloshuk and J.W. Bennett (2004) Clustered pathway genes in aflatoxin biosynthesis. *Appl Environ Microbiol*, **70**, 1253–1262.

W.J. Zhang, C. Pedersen, M. Kwaaitaal, P.L. Gregersen, S.M. Mørch, S. Hanisch, A. Kristensen, A.T. Fuglsang, D.B. Collinge and H. Thordal-Christensen (2012) Interaction of barley powdery mildew effector candidate CSEP0055 with the defence protein PR17c. *Mol Plant Pathol*.

X. Zhang, C. Scheuring, S. Tripathy, Z. Xu, C. Wu, A. Ko, S.K. Tian, F. Arredondo, M.K. Lee, F.A. Santos, R.H.Y. Jiang, H.B. Zhang and B.M. Tyler (2006) An integrated BAC and genome sequence physical map of *Phytophthora sojae*. *Mol Plant Microbe Interact*, **19**, 1302–1310.

Z. Zhang, C. Henderson, E. Perfect, T.L.W. Carver, B.J. Thomas, P. Skamnioti and S.J. Gurr (2005) Of genes and genomes, needles and haystacks: *Blumeria graminis* and functionality. *Mol Plant Pathol*, **6**, 561–575.

H. Zhu, S. Choi, A.K. Johnston, R.A. Wing and R.A. Dean (1997) A large-insert (130 kbp) bacterial artificial chromosome library of the rice blast fungus *Magnaporthe grisea*: genome analysis, contig assembly, and gene cloning. *Fungal Genet Biol*, **21**, 337–347.

D. Zohary, M. Hopf and E. Weiss (2012) *Domestication of plants in the old world*. Oxford University Press.

Acknowledgements

I would like to thank...

...Beat for giving me the chance to work on this (a bit unusual PhD-) project, which added a few more shades of grey to my hair and made me loose some illusions, but also allowed me learn lots of lessons, visit inspiring conferences and be part of stimulating collaborations.
...Thomas for being a very patient supervisor and a great person to work with, and for motivating me with his optimism and enthusiasm for Science.
Röbi for being in my PhD committee, great story-telling at apéros, discussions about career planning at coffee breaks and for reminding me of the importance of common sense.
...the mildew people Francis, Roi, Stefan, Salim and Kaitlin for their team spirit, enthusiasm and impressive commitment, and for being "fun guys" even if work is sometimes not much fun. Merci á Francis for all the DNA extractions, good fun during our "honeymoons", great teamwork when we were still lonesome cowboys on the project and his awesome humour.
...the current and former members of lab P3/12 James, Jan, Stefan, Kostas, Thomas, Chris and Margarita for the nice working-atmosphere, answers to stupid questions, lame jokes, the maxibar and lots of fun at and after work - thank you for nothing! Special thanks go to Jan for helping me with computer- and programming problems and sharing the same taste of music, and to Kostas for unforgettable holidays, road trips and outings.
...the "ladies" Lisi, Margarita, Elena and Jyoti for sharing gossip, nervous breakdowns, baby experiences and the passion for the opera, for Wodka shots on Friday evenings, hat sessions at weddings, funny food in Beijing's side streets and discussions about style, love and life. You made the grey lab shine in colours.
...Matthias, Draeger, Beni, Christina, Gabi, Aurel, Mayank and everybody else from the institute who shared with me the daily ups and downs of a PhD, fantastic road trips and cold beers on the roof.
...the members of the Keller- and Dudler group for the nice atmosphere, valuable discussions and technical support (thank you Gabi and Geri!).
...all the people who contributed to the development of open-source software I used for this project.
...my friends Katja, Andrea, Ariane, Cynthia, Katharina, Souria, Sonia, Olivia, Dodo, Stefi, the "Bio-Mädels" and the members of "Club Français" for their support and friendship, and my basketball team for making Monday my favourite day of the week.
...Christof, who again and again found the right words to put me back on my feet when I could not see the end of the tunnel. Thank you so much for your patience, your interest in my work and your love.
...my family, and in particular my parents, who have always encouraged and supported me.

Curriculum Vitae

Surname:	OBERHÄNSLI
Name:	Simone
Date of birth:	31.10.1980
Place of origin:	Kemmental TG

Education

1996-2001	Lehramt, Matura Typ L Kantonsschule Stadelhofen
2003-2006	Undergraduate studies in Biology, University of Zurich
2006-2007	MSc thesis under supervision of Prof. Beat Keller, Institute of Plant Biology, University of Zurich Title: Nuclear localization of PM3 and development of molecular tools for genetic studies on wheat powdery mildew.
December 2007:	MSc in Biology, Plant Sciences, University of Zurich
July 2008-present:	PhD thesis under supervision of Prof. B.Keller at Institute of Plant Biology, University of Zurich

APPENDIX A

Supplementary Material Chapter 2

A.1 Supplementary Tables

Table A.1. Overlapping BAC clones for two genomic regions. PCR screening of the library was done on plates 1 to 16. Molecular markers used to screen for locus 2 have been described in Oberhaensli *et al.* (2011). Screening for locus GTCA_E4 was done with molecular marker GTCA_E4 (Parlange and Keller, unpublished data)

	Clones screened by PCR	Presence in the FPC assembly
Locus1	2P10	yes
	6M19	yes
	9N07	no[a]
	15H07	yes
	15K07	yes
	6C19	yes
Locus GTCA_E4	2H08	yes
	7H01	yes
	8O11	no[a]
	14P08	yes
	15G21	yes

[a] Those BACs were missing from the 6,831 BACs used for physical map assembly

A.2 Supplementary Figures

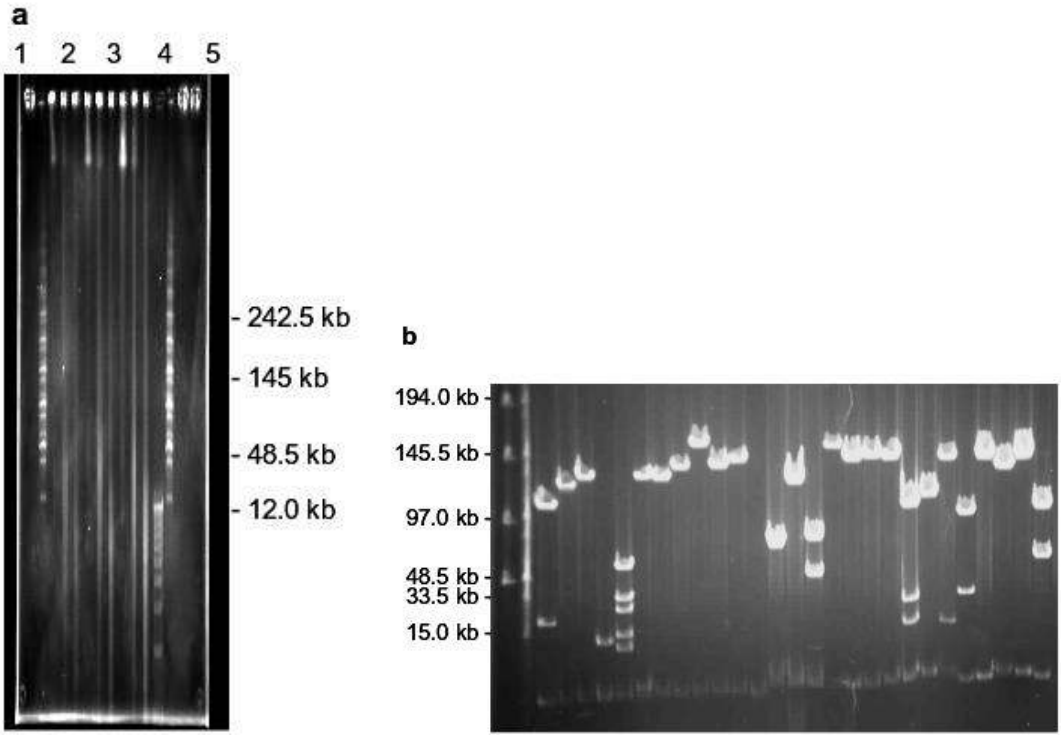

Figure A.1. Construction of a BAC library from *B.g. tritici* DNA. **a.** Digestibility tests of HMW DNA. DNA was tested for digestibility using restriction enzyme *Hind*III. Lane 1: undigested control; lane 2: partial digestion with 10U/ml for 20 minutes; lane 3: complete digestion with 100U/ml for 6 hours; lane 4: 1kb ladder; lane 5: MidRange PFG Marker I. **b.** Insert size analysis of 27 randomly selected BAC clones. BAC DNA was digested with *Not*I and separated by PFGE. Markers on the left are Lambda Ladder PFG Marker and MidRange Marker I.

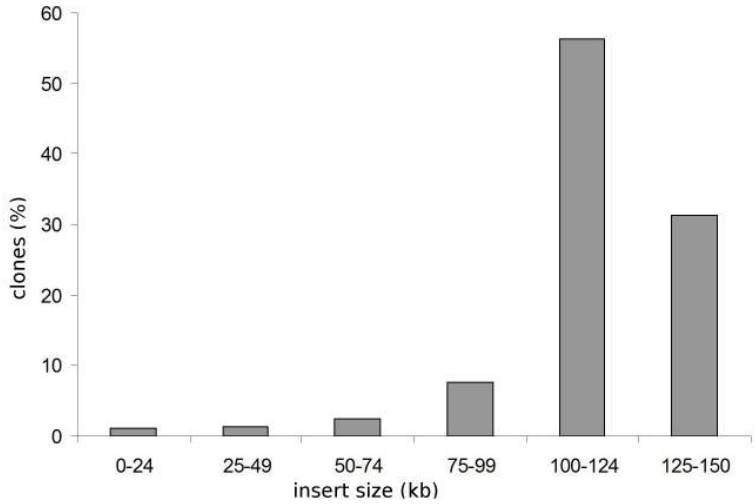

Figure A.2. Insert size distribution in the BAC library. Insert sizes were analysed in 300 BAC clones randomly selected from the three fractions of the library (B, M1, M2). The overall distribution of insert sizes was calculated by combining data from the three fractions considering their proportion in the library.

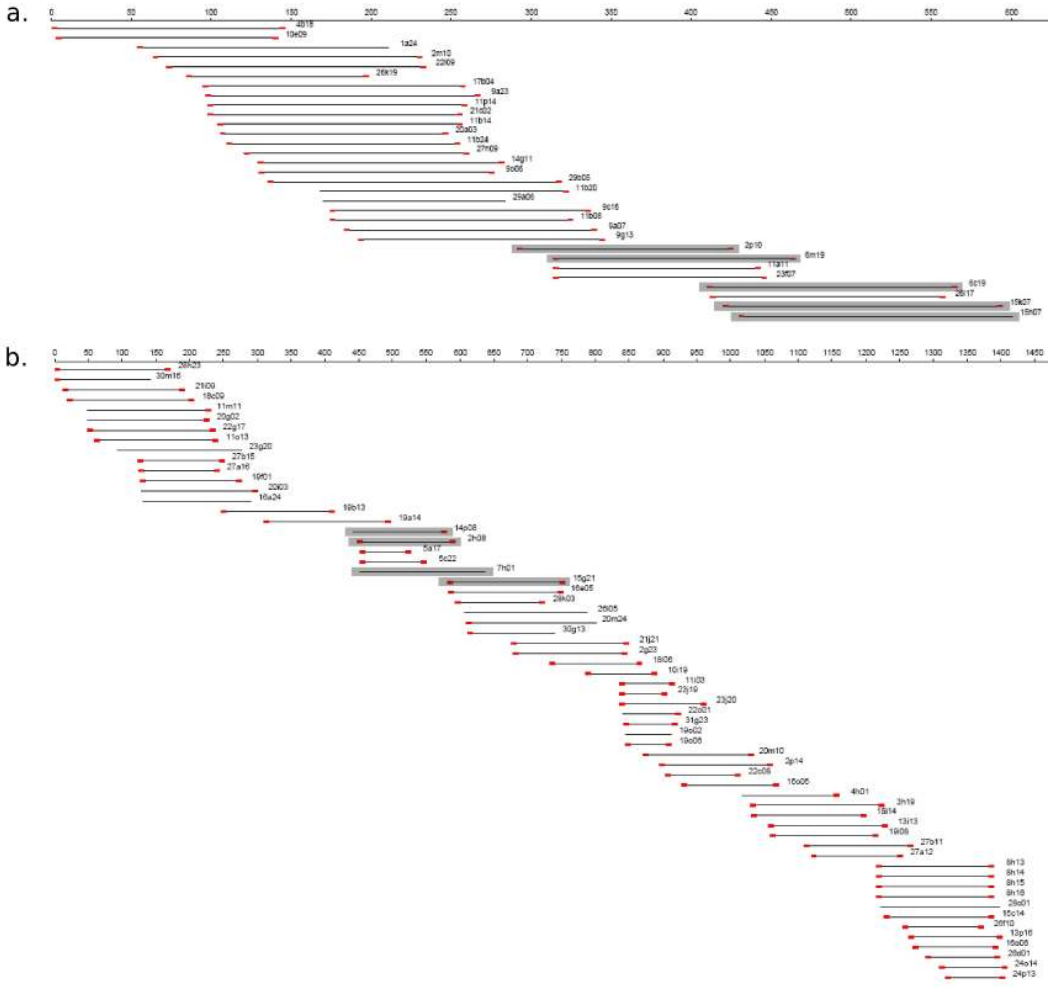

Figure A.3. BAC contigs comprising locus 2 and locus GTCA_E4. **a.** Contig ctg5, harbouring the locus 2 (Oberhaensli *et al.*, 2011). **b.** Contig ctg25, harbouring the locus GTCA_E4. Scale is in kilobases. BAC clones identified by PCR-screening of the library are highlighted in gray. Red boxes at the BAC-ends indicate the availability of the respective BES (note that the orientation is unknown). The graphical representation was produced based on the FPC files and using WICKERsoft software.

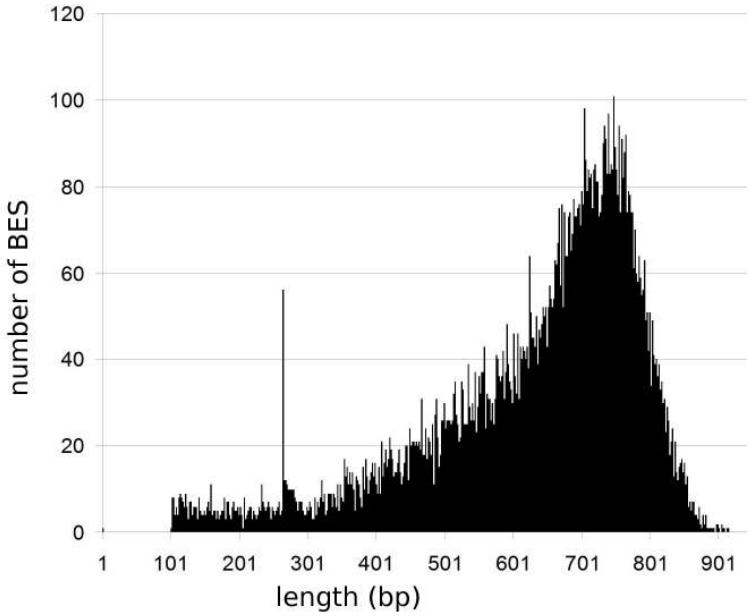

Figure A.4. Size distribution of BAC-end sequence (BES) length. A total of 20,001 BES were generated by the sequencing from both ends of the entire BAC library. The average read length is 633 bp with 82% of the reads being above 500 bp. The peak observed at 263 bp is caused by 54 identical sequences which were shown by BLAST analyses to correspond to the sequence of the origin of replication "transposable R6K ori" (Stojiljkovic *et al.*, 1994).

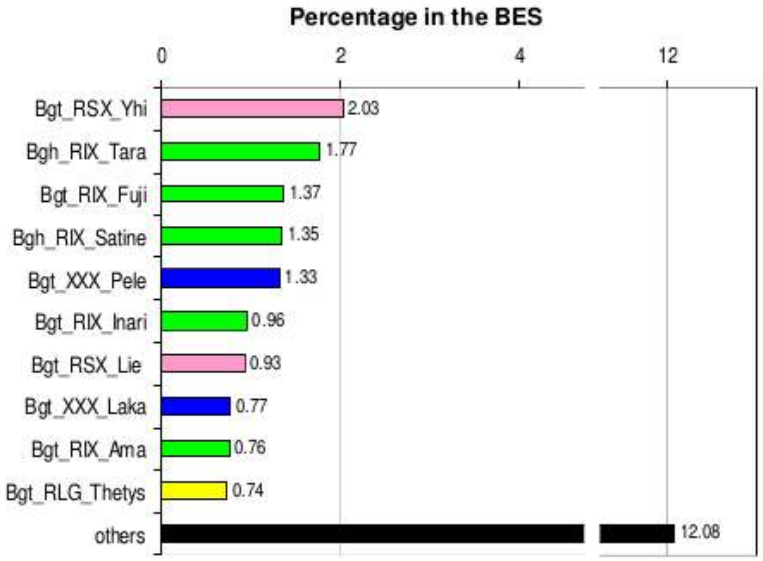

Figure A.5. Distribution of the ten most abundant TE families of the *Blumeria* Repeat Database in the BES. Bgt and Bgh indicate origin of TE (Bg tritici and Bg hordei, respectively). Names are according to the nomenclature of (Wicker *et al.*, 2007): RSX, SINE (pink); RIX, LINE (green); RLG, Gypsy (yellow); XXX, unclassified (blue).

APPENDIX B

Supplementary Material Chapter 3

B.1 Supplementary Tables

Table B.1. Conservation of *B.g. tritici* and *B.g. hordei* gene coding sequences (CDS)

	Locus1		Locus2	
Gene number	Identity CDS	Identity incl. introns	Identity CDS	Identity incl. introns
1	95.80%	96.0%	89.8%	89.8%
2	94.2%	94.4%	92.5%	91.5%
3	92.1%	91.6%	86.4%	no intron
4	94.3%	94.0%	93.5%	92.4%
5	87.5%	87.1%	93.1%	93.1%
6	81.0%	no intron	94.7%	94.4%

B.2 Supplementary Figures

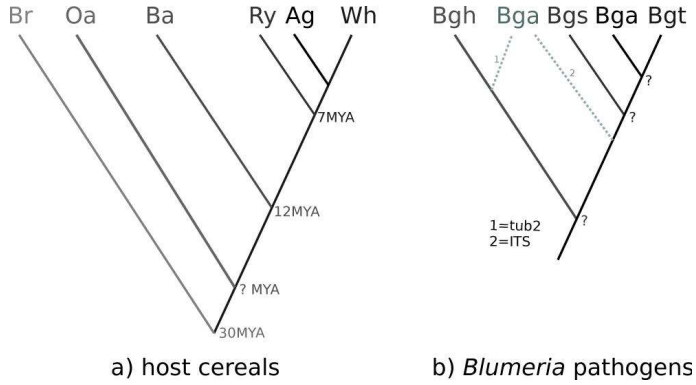

Figure B.1. Phylogenetic relationship of host cereals including Brachypodium (Chalupska *et al.*, 2008; Draper *et al.*, 2001; Initiative, 2010) and the corresponding *Blumeria* ff. spp. Abbreviations: *Brachypodium distachyon* (Br), *Avena sativa* (Oa), *Hordeum vulgare* (Ba), *Secale cereale* (Ry), *Agropyron christatum* (Ag), *Triticum aestivum* (Wh), *B.g. hordei* (Bgh), *avenae* (Bga), *secali* (Bgs), *agropyri* (Bgg) and *tritici* (Bgt). The phylogenetic position of Bga was determined by Wyand and Brown (2003) based on rDNA internal transcribed spacer (ITS) and β-tubulin (tub2) gene

APPENDIX C

Supplementary Material Chapter 4

C.1 Supplementary Tables

Table C.1. *B.g. tritici* isolate 96224 genome reference sequence statistics

454 assembly statistics		
Total read number	11,330,743[a]	
Assembled reads	9,846,293 (86.8%)	
Sequence coverage	13x	
Scaffolds	3,522	
Size	97.4 Mb	
Total N's[b]	14 Mb	
Scaffold N50[c]	48.7 kb	
Largest scaffold	290 kb	
Reference genome statistics	anchored[d]	total[e]
FP contigs	250	251
Size (incl. N's)	107 Mb	126 Mb
Number of non-N bases	67 Mb	82 Mb
Average size of a non-anchored 454 scaffold	11.5 kb	

[a] Assembly includes whole genome shotgun sequences (454) and 20'000 BAC end sequences (Sanger)
[b] N bases represent sequence gaps
[c] Half of the assembly consists of scaffolds larger than this size
[d] 454 scaffolds anchored on contigs from a BAC library finger print (FP) assembly
[e] 454 scaffolds that could not be anchored were collected in an extra pseudomolecule (Bgt_ctg-10000)

Table C.2. Characterisation of *B.g. tritici* genes. The characterisation was done in hierarchical order. First, all genes were tested for homologs in yeast. Those which did not have a homolog were then used as query against PFAM, followed by blasts against the genomes of *Botrytis cinerea* and *B.g. hordei*.

Characterisation	Genes
Homolog in yeast	3,182
Homolog to PFAM domain	205
Homolog in *Botrytis cinerea*	1,344
Homolog in *B.g. hordei* [a]	1,442
CSEP[b]	437
CEP[c]	165
Uncharacterised genes	367

[a] Includes [b] and [c].
[b] Candidate secreted effector protein, identified through homology to *B.g. hordei* CSEPs.
[c] Novel candidate effector proteins without a signal peptide.

Table C.3. Sizes of the largest gene families encoding candidate effector proteins (CEPs)

Gene family	total[a]	*B.g. tritici* [b]	*B.g. hordei* [c]
FAM_E0014-134	14	12	2
FAM_E0014-262	14	2	12
FAM_E0016-115	16	6	10
FAM_E0019-107	19	13	6
FAM_E0020-1	20	12	8
FAM_E0046-114	46	25	21
FAM_E0067-10	67	31	36
FAM_E0200-102	200	124	76

[a] Total number of gene in family.
[b] Number of *B.g. tritici* genes in family.
[c] Number of *B.g. hordei* genes in family.

Table C.4. Presence/absence polymorphism of genes in nine *B.g.tritici* isolates compared to the reference isolate 96224. Presence and absence of a gene is indicated with + and -, respectively. CSEP: candidate secreted effector protein, CEP: candidate effector protein, P: gene is present but open reading frame is interrupted either by a frameshift or a premature stopcodon, (-): gene is partially deleted

Gene	215[a]	8[a]	97[a]	15[a]	217[a]	70[ad]	7004[b]	94202[bd]	JIW2[cd]	Gene product
BgtE-5545	+	+	+	+	+	+	-	+	-	CSEP
BgtE-5597	-	-	-	-	-	-	-	-	-	CSEP
BgtE-5802	-	-	-	-	-	-	-	-	-	CSEP
BgtE-5845	+	+	-	+	+	+	-	+	-	CSEP
BgtE-5419	+	+	P	+	P	P	+	-	+	CSEP
BgtE-3419	-	+	+	-	+	-	+	+	+	CSEP
BgtAc-30466	P	+	+	+	+	+	+	-	+	CSEP
BgtAc-31249	+	+	(-)	+	+	+	+	-	+	CSEP
BgtAcSP-30824	+	-	+	-	+	-	+	+	+	CSEP
BgtE-40100	+	-	+	+	+	+	-	+	-	CSEP
BgtA-21525	-	-	-	+	+	-	+	+	-	CEP
Bgt-4055	-	-	-	+	+	-	-	+	+	CEP
BgtA-20784	+	+	+	+	+	-	+	+	+	CEP
Bgt-369	+	+	-	+	+	+	+	+	-	other
BgtAc-31336	-	-	-	+	-	-	+	-	+	no homolog
BgtA-20381	-	-	-	+	+	+	-	-	-	no homolog

[a] Country of origin Israel
[b] Country of origin Switzerland
[c] Country of origin UK
[d] Genome analysed in this study

Table C.5. Numbers of polymorphisms between three *B.g. tritici* isolates and the *B.g. tritici* reference genome

Type of polymorphism	70	94202	JIW2
Total SNPs[a]	233,997	175,093	182,904
High confidence SNPs[b]	161,117	116,687	113,967
Ti/Tv[c]	1.51	1.52	1.49
SNPs in CDS[d]	6,289	4,389	4,514
Non-syn[e]	2,815	1,982	2,039
Total insertions	5,253	4,616	3,858
Total deletions	4,545	3,390	3,022
High confidence InDels[f]	1,797/1,981	1,322/1,365	1,447/1,360

[a] Total number of sites that showed base substitutions
[b] For definition of high confidence SNPs see suppl. text 4.3.7
[c] Ratio of transitions to transversions
[d] Number of SNPs in coding sequences of genes
[e] Number of SNPs that cause amino acid changes in genes
[f] For definition of high confidence insertions/deletions see suppl. text 4.3.7

Table C.6. Numbers of different protein variants in the four *B.g. tritici* isolates studied. As protein variants was called if the predicted protein sequence of two isolates differed

Number of protein variants[a]	All genes	non-CEPs[b]	CEPs[c]
1	3,684 (56.8%)	3,413	271
2	1,969 (30.4%)	1,761	208
3	692 (10.7%)	597	95
4	141 (2.2%)	123	18

[a] Number of different protein variants per gene
[b] All genes except CSEPs and CEPs
[c] Includes CSEPs and CEPs

Table C.7. Pairwise comparisons of fractions of the genome where two *B.g.tritici* isolates share the same haplogroup. Indicated is for each pairwise comparison the number of genomic segments where the two isolates are of the same haplogroup. The number in parentheses is the fraction of the genome that is of the same haplogroup.

	JIW2	94202	70
96224	618 (25.1%)	547 (26.3%)	321 (6.9%)
JIW2	-	622 (24.2%)	366 (8.0%)
94202	-	-	394 (8.6%)

Table C.8. Divergence time estimates for different haplogroups in the *B.g. tritici* genome.

Isolate	Haplogroup	Fraction[a]	SNPs/kb[b]	Divergence[y]	SD[c]
JIW2	H_{young}	25.1%	0.11	5,708	±3,087
94202	H_{young}	26.3%	0.11	5,407	±3,241
70	H_{young}	6.9%	0.22	8,690	±3,054
JIW2	H_{old}	74.9%	1.11	55,423	±12,156
94202	H_{old}	73.3%	1.20	60,439	±13,374
70	H_{old}	93.1%	1.31	63,157	±13,222

[a] Contribution of the respective haplogroup to the genome
[b] Average SNP density for the haplogroup
[c] Standard deviation

Table C.9. PFAM domains present in obligate biotrophic and non-obligate biotrophic fungi.

Species	Total genes	PFAM hits[a]	PFAM families[b]
B.g. tritici	6,540	2,010	1,182
A. nidulans	10,560	3,382	1,439
M. grisea	11,054	3,110	1,382
B. cinerea	16,448	3,059	1,329
P. graminis	20,565	2,459	1,079

[a] Notal number of genes with homology to PFAM domains
[b] Number of different PFAM domains identified in the genome

Table C.10. *De novo* assembly statistics of isolates 94202, JIW2 and 70.

	94202	JIW2	70
nr of reads (assembled)	56,838,638	15,208,328	54,776,180
nr of contigs	94,070	86,131	93,060
assembly size	72 Mb	65 Mb	77 Mb
average sequence contig size	767 bp	762 bp	831 bp
largest sequence contig	29,239 bp	23,215 bp	31,455 bp

C.2 Supplementary Figures

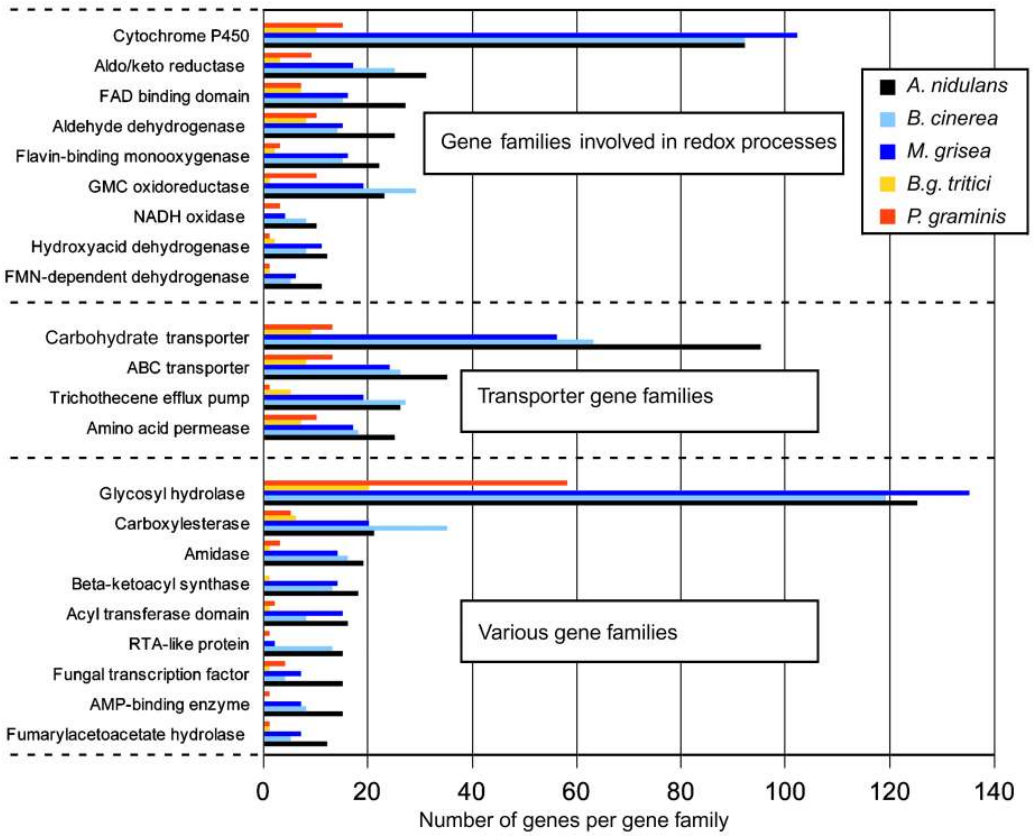

Figure C.1. Gene families which are reduced in size in the genomes of obligate biotrophic fungi compared to the genomes of heterotrophic fungi. Horizontal bars indicate the number of genes belonging to particular gene families. The fungal species *A. nidulans*, *B. cinerea* and *M. grisea* have at least some heterotrophic growth phase during their life cycle while *B.g. tritici* and *P. graminis* are obligate biotrophs. AMP: Adenosine mononucleotide, FAD: Flavin adenine dinucleotide, FMN: Flavin mononucleotide, GMC: Glucose-methanol-choline, NADH: Nicotinamide adenine dinucleotide, RTA: R transactivator.

Figure C.2. Glycosyl hydrolases from *A. nidulans*, *B. cinerea*, *M. grisea*, *B.g. tritici* and *P. graminis*. The columns at the left show the number of genes in each family for the 5 species. Colour-coding indicates whether families are completely absent and in which species that is the case.

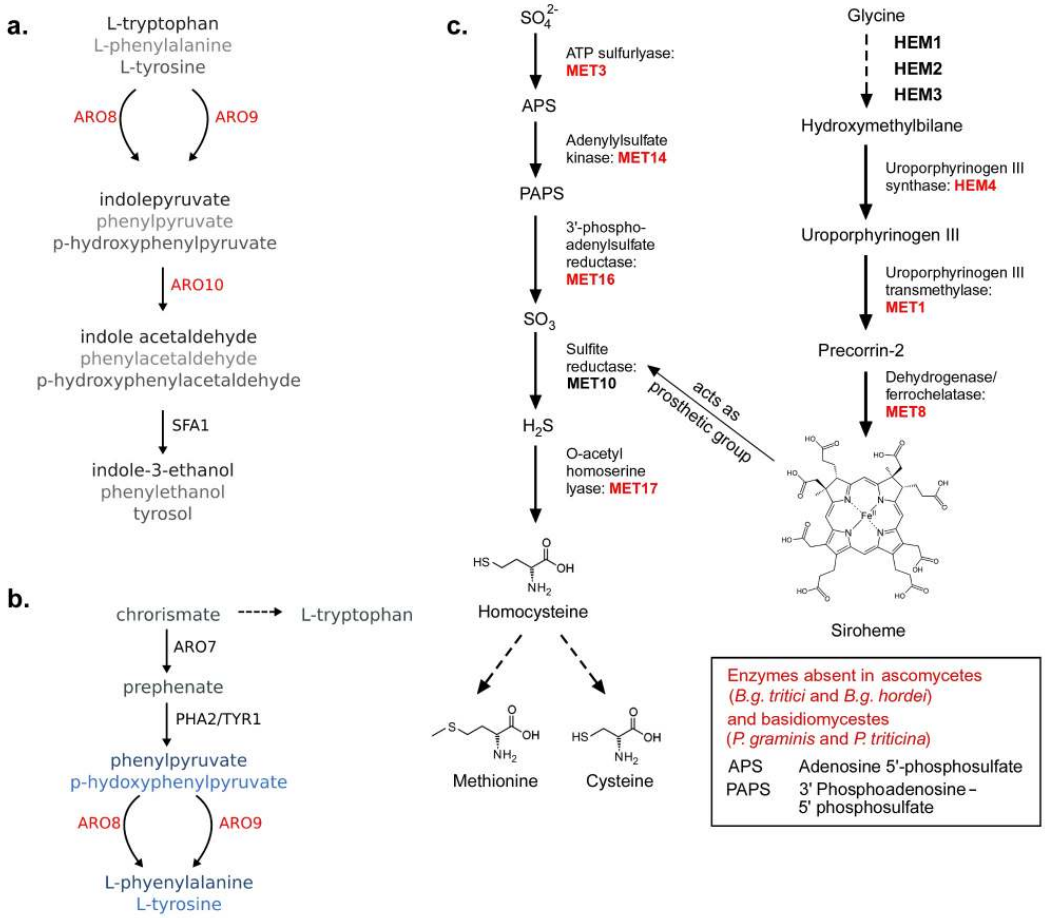

Figure C.3. Enzyme deficiencies in the amino acid metabolism of obligate biotrophic fungi. Enzymes which are absent from the genomes of powdery mildew and rust fungi are depicted in red. Considering their large phylogenetic distance, powdery mildews and rusts must have lost the exact same sets of enzymes independently. **a.** Catabolism pathways of tryptophane, phenylalanine and tyrosine are missing the enzymes which convert the amino acid backbone to acetaldehyde. **b.** Two of the enzymes described in **a.** (ARO8 and ARO9) are also essential for the final step of the phenylalanine and tyrosine biosynthesis. **c.** Deficiencies in the assimilation of inorganic sulfur in the biosynthesis of methionine and cysteine. Both the direct pathway of sulfur assimilation as well as the pathway which produces a prosthetic group for MET10 are absent.

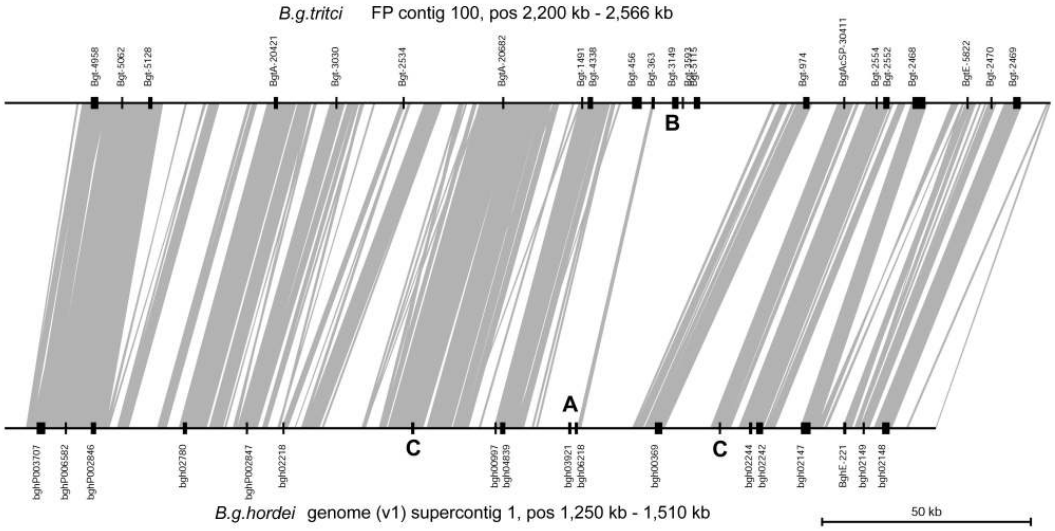

Figure C.4. Comparison of approx. 300 kb of genomic sequences from *B.g. tritici* and *B.g. hordei*. Sequences which are conserved between the two *formae speciales* are connected with grey areas. Genes are indicated as black boxes. In one region, gene content differs between the two *formae speciales* (indicated with A and B). This is likely due to misassemblies as all four *B.g. tritici* genes that are absent from the *B.g. hordei* sequence are found elsewhere in the *B.g. hordei* genome. However, the *B.g. hordei* bgh03921 gene (indicated with A) has no homolog in the *B.g. tritici* genome sequence and might therefore represent a true presence/absence difference. Two previously un-annotated genes in *B.g. hordei* were identfied during the comparative analysis (indicated with C).

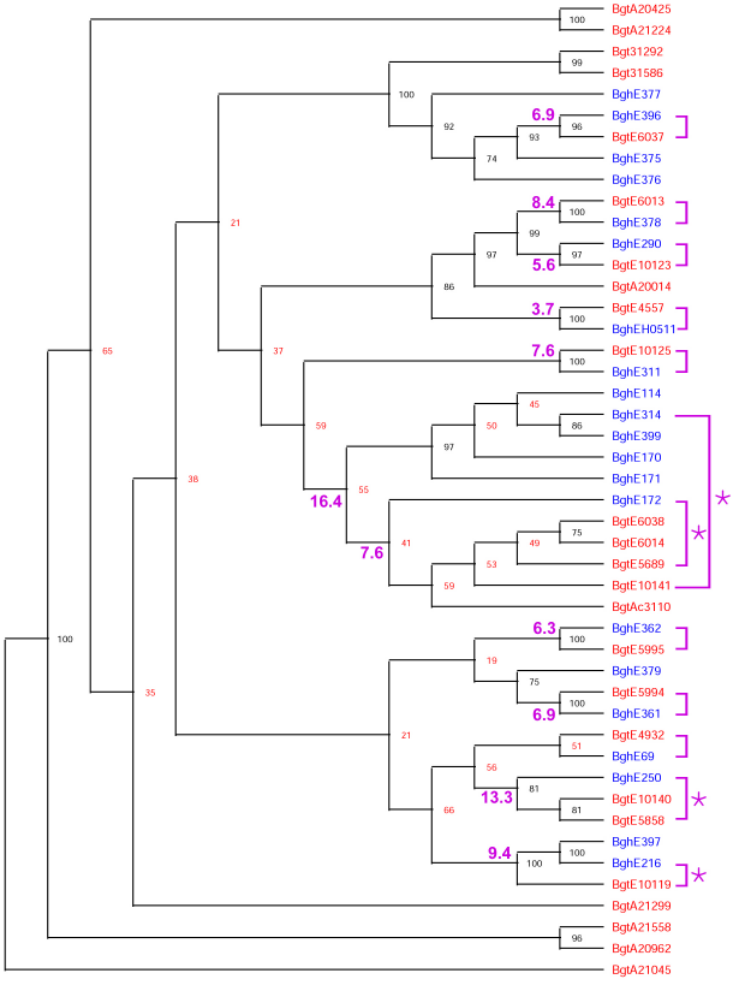

Figure C.5. Phylogenetic tree of CSEP family FAM46-114. *B.g. tritici* genes are displayed in red while *B.g. hordei* genes are displayed in blue. The tree is a consensus tree of 100 bootstrap replica. The numbers at forks indicate the number of times the group consisting of the sequences which are to the right of that fork occurred among the trees, out of 100 trees. Bi-directional closest homologs (BDCHs) of *B.g. tritici*/*B.g. hordei* genes (determined by blast searches) are indicated by purple brackets. Those BDCHs which lie on distantly related branches are considered deep paralogs (indicated with asteriks). Divergence time estimates for the BDCH pairs are indicated as purple numbers at the respective branchings of the tree. Divergence time estimates have error rates of $\pm$23-32%, depending on the length of the proteins

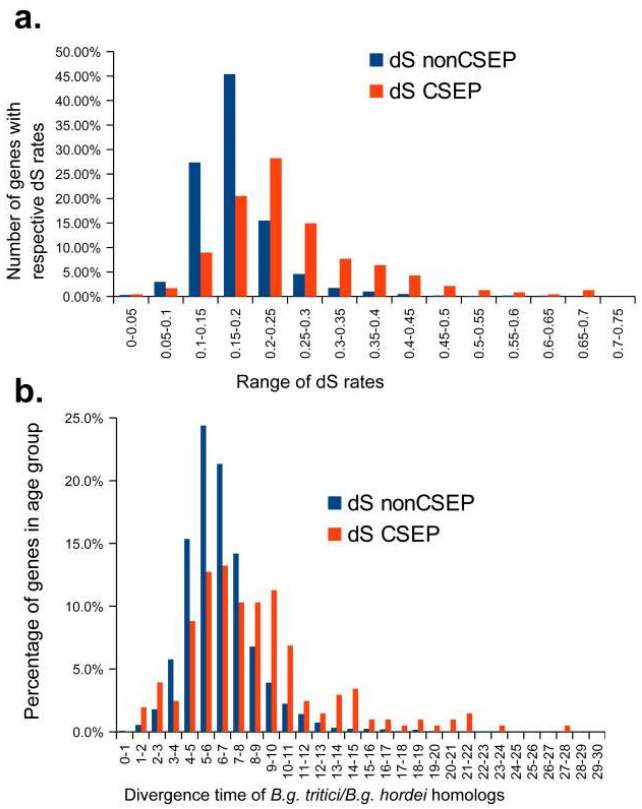

Figure C.6. Estimation of the number of deep paralogs in CSEPs. **a.** Distribution of rates of synonymous substitutions per synonymous sites (dS) in 5,258 closest *B.g. tritici*/*B.g. hordei* gene homologs. Rates were calculated with the yn00 program of the PAML software package. Displayed are the distributions of dS rates of 5,021 non-CSEP and 237 CSEP gene pairs. The x-axis shows the range of dS rates in the two gene groups while the y-axis shows the number of gene pairs in each range. The y-axis numbers are given in % of the total number of gene pairs in each group (i.e. 5,021 non-CSEP and 237 CSEP). **b.** Distribution of divergence time estimates of non-CSEPs and CSEPs. Using the non-CSEPs as a reference, one can infer that CSEPs age groups whose sizes deviate strongly from those in the non-CSEPs contain deep paralogs. For example, one can assume that most *B.g. tritici*/*B.g. hordei* CSEP pairs which diverged more than 10 Myr ago are deep paralogs. Using the cutoff of 10 Myr, at least 20% of CSEP pairs represent deep paralogs..

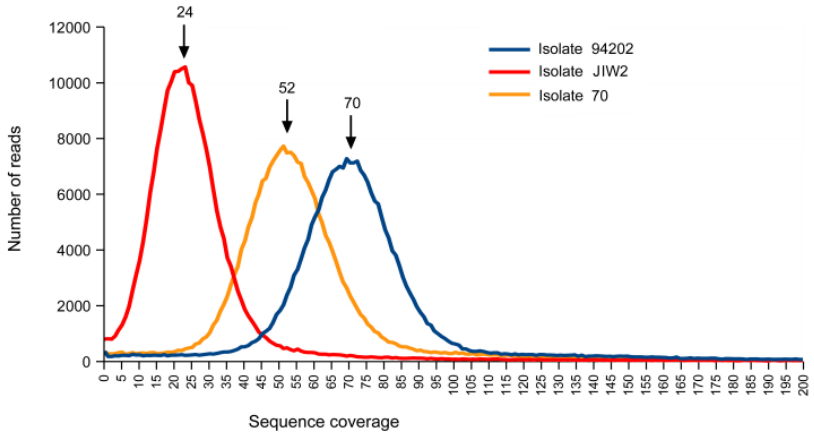

Figure C.7. Sequence coverage of three *B.g. tritici* isolates with Illumina reads. The x-axis is the sequence coverage. The y-axis indicates the number sequence reads within each coverage range. The coverage plots were derived from 500'000 random positions in the genome. Numbers above arrows indicate the average coverage of single-copy sequences ("peak coverage").

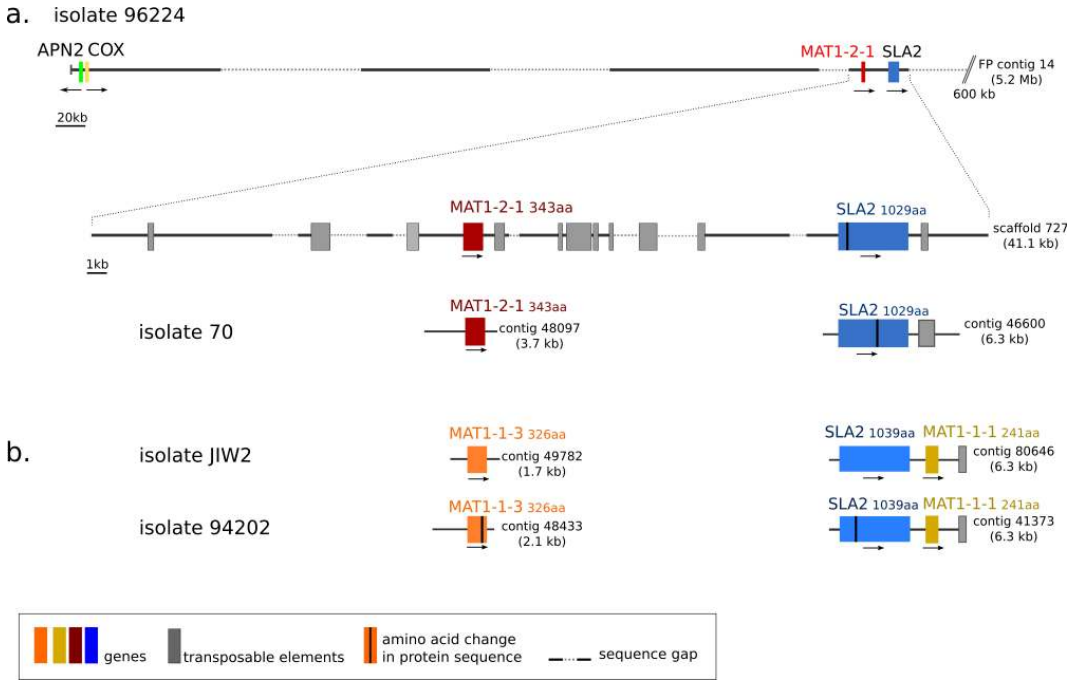

Figure C.8. Schematic representation of the mating type locus in *B.g. tritici*. **a.** Isolate 96224 and isolate 70 have a *MAT1-2* idiomorph. The scheme describes the position of the *MAT1-2-1* gene and the flanking genes *SLA2* and *APN2/COX* on FP contig 14 of the reference genome (isolate 96224). Arrows indicate the orientation of transcription. A close-up of the MAT locus reveals that *MAT1-2-1* and *SLA2* are located relatively close to each other and are surrounded by transposable elements. The MAT1-2 protein sequences of isolate 70 and 96224 share 100% identity, while the *SLA2* proteins differ in two amino acids (black bar). **b.** Isolates JIW2 and 94202 have a *MAT1-1* idiomorph. *MAT1-1-3* and *MAT1-1-1* are located on different contigs of the Illumina *de novo* assemblies. The *SLA2* gene is located immediately upstream of the *MAT1-1-1* coding region. The *MAT1-1-1* proteins are conserved, whereas the *MAT1-1-3* and the *SLA2* proteins differ in one amino acid between the two isolates.

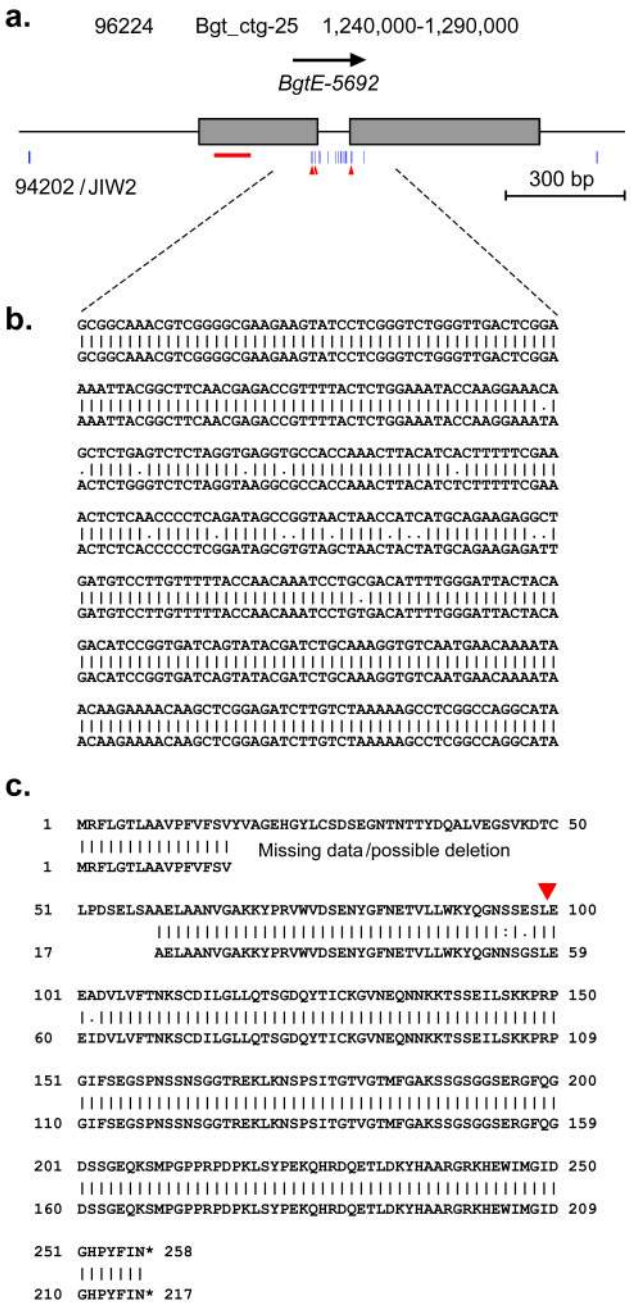

Figure C.9. **a.** The candidate effector gene *BgtE-5692* contains a highly divergent segment covering parts of exons 1 and 2 (grey boxes) as well as the intron. The small size of the divergent fragments suggests that it was introduced through gene conversion. SNPs are indicated as blue vertical bars. Three SNPs which cause changes in the encoded protein are indicated with red arrowheads. The red bar indicates a 110 bp region which was not covered by the Illumina assembly in both JIW2 and 94202. **b.** DNA sequence alignment of the variable region plus flanking sequences. **c.** Alignment of the predicted protein sequences showing the amino acid changes. The position of the intron is indicated with a red triangle. The sequence gap in the N-terminal region (missing data) corresponds to the sequence gap indicated with the red bar in **a.**.

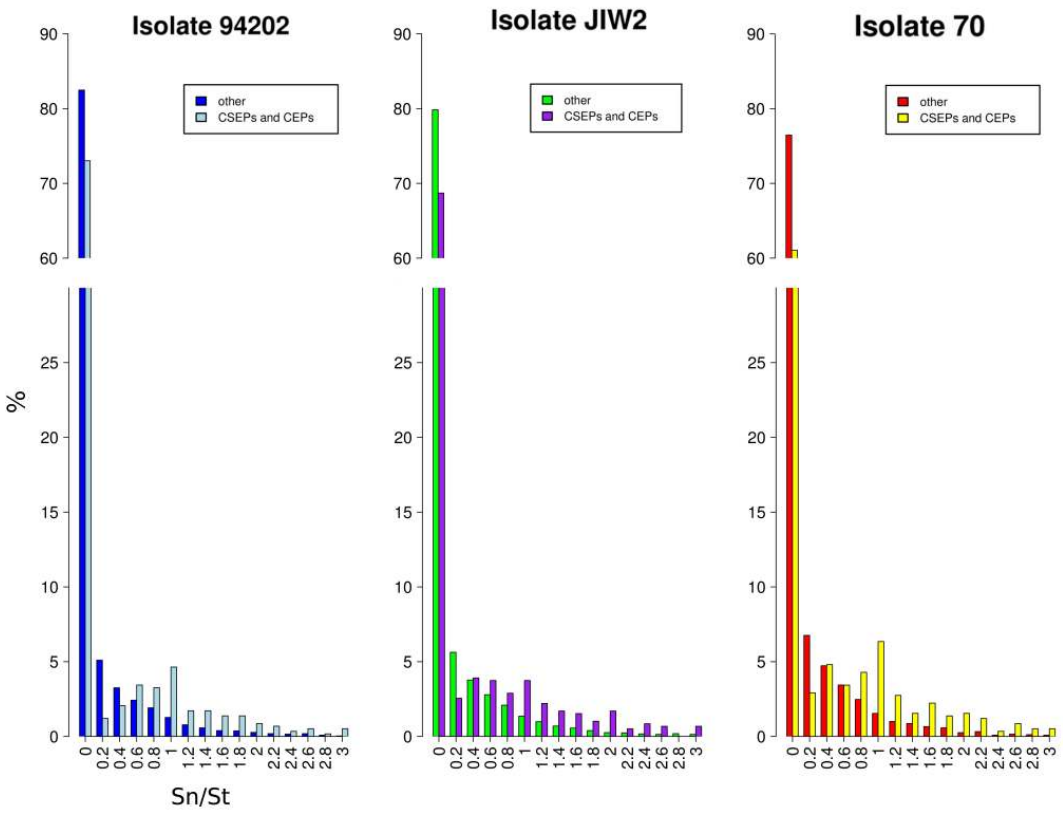

Figure C.10. Non-synonymous base exchanges are more frequent in CSEP and CEP genes. For each of the isolates JIW2, 94202 and 70, coding sequences of genes were aligned with their ortholog from reference isolate 96224. For each gene, the ratio of non-synonymous substitutions compared to the total number of substitutions was calculated. The value was normalized for gene size (i.e substitutions per kb). The dataset obtained for each isolate was split in CSEPs/CEPs and non-CSEPs/CEPs, respectively. The values for the genes of each group were classified in bins of 0.2, e.g. bin 1 contains all genes with values from 0 to 0.2, bin 2 contains values from 0.2 to 0.4. After counting the number of genes per bin, the respective percentage on the total number of genes of the group was plotted in a bar-diagram. Sn=number of non-synonymous base exchanges, St=total number of base exchanges.

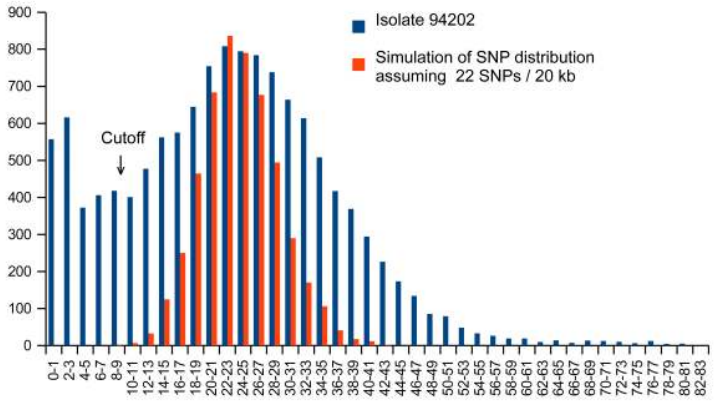

Figure C.11. Characterisation of haplogroups in wheat powdery mildew. SNP density in the genome was surveyed in windows of 20 kb with a 2 kb sliding step. The x-axis indicates the number of SNPs per 20 kb window while the y-axis indicates how many 20 kb windows contained the respectiuve number of SNPs. The blue series shows SNP distribution of isolate 94202. The genome is composed of segments with high and low SNP densities, which are reflected in the two peaks of the curve. Segments with a higher SNP density produce a peak at 22 SNPs. The red series shows the distribution of SNP densities, assuming a purely random SNP distribution and an average SNP density of 22 SNPs per 20 kb. This simulation was used to determine the cutoff value (9SNPs/20 kb) to distinguish SNP-rich from SNP-poor regions.

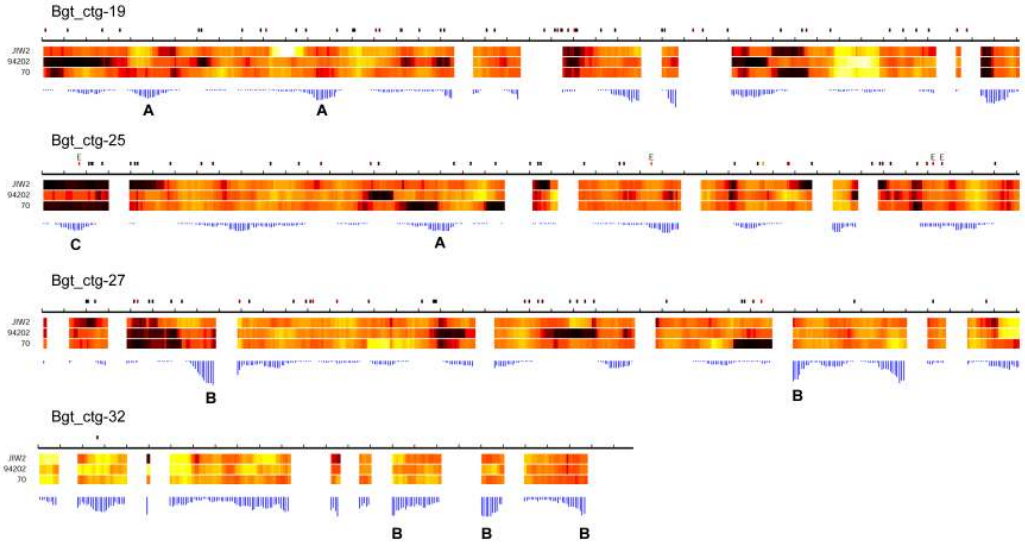

Figure C.12. Quality control of SNPs in three *B.g. tritici* isolates. The heatmap displays SNP densities in 10 kb sliding windows with a 1 kb sliding step. Gaps smaller than 1 kb were ignored for the calculation of SNP frequencies. White spaces display positions of major sequence gaps (< 1kb) where no sliding windows could be made across. Warmer colours indicate higher SNP frequencies. The blue bars underneath the map indicate the amount of discarded SNPs in each window (i.e. SNPs with low quality). Dots above the heatmap represent genes, CSEPs/CEPs are marked with "E". Regions where some proportion of SNPs had to be discarded occur frequently both in overall SNP-rich (examples A) and SNP poor regions (example C). However, the discarded SNPs are usually a minority of the total SNPs in a region. Only in extremely repetitive regions (examples B), the number of discarded SNPs was considerable. The main concern was that regions with many discarded SNPs would be mistaken for SNP-poor regions. This was shown not to be the case.

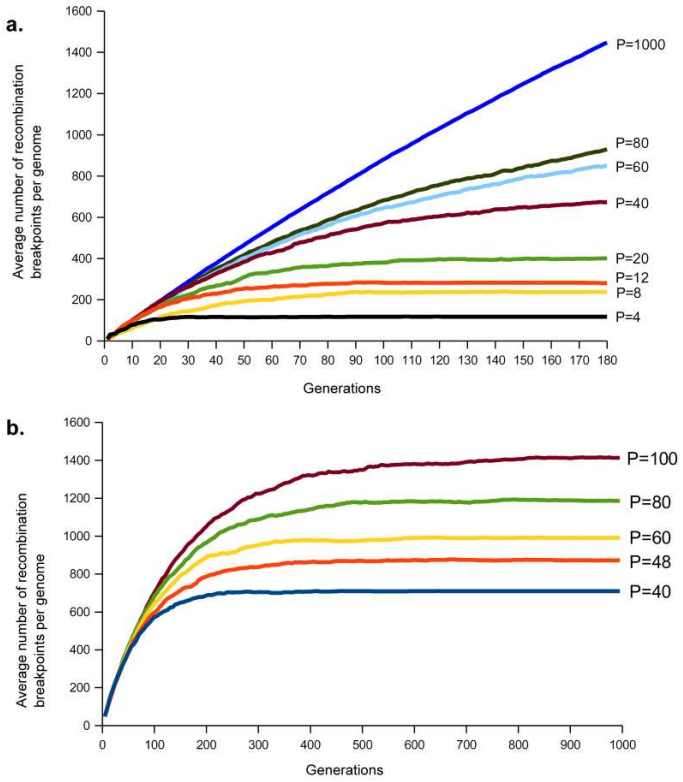

Figure C.13. Simulations of the number of recombination breakpoints in genomes of haploid species with two mating types without mating restrictions. The x-axis is the number of generations while the y-axis indicates the average number of haplogroup breakpoints that are expected per individual. A genetic map of the size 2,000 cM was assumed. Furthermore, we assumed that the starting population consists of two sub-populations of two different haplotypes. The size of the population is indicated with P next to the corresponding graphs. **a.** At large population sizes (e.g. 1,000), the number of haplogroup breakpoints increases in a linear way while for smaller populations, the curves rapidly converge toward a stable value. **b.** In populations of 40-100 individuals, the numbers of haplogroup breakpoints converge rapidly to values similar to those observed in the studied *B.g. tritici* isolates.